Measuring disability in neurological rehabilitation: psychometric evaluation of two outcome measures

Thesis submitted to the University of London

for the degree of Doctor of Philosophy, September 1999

Jeremy Charles Hobart

Health Services Research Unit

Department of Public Health and Policy

London School of Hygiene & Tropical Medicine

Keppel Street, London WC1E 7HT, UK

For Claire.....who waited (patiently)

Acknowledgements

This project could not have been accomplished without the contribution and help of many people. To these people, I am deeply grateful.

First, my sincere thanks go to my supervisor Dr. Donna Lamping. Her clarity of thought, attention to detail, and commitment to excellence in scientific enquiry is second to none. She has helped me enormously in the development of my own expertise in health measurement.

I am indebted to Professor Alan Thompson at the Institute of Neurology, whose expertise, wisdom, personal sacrifices, motivation, and friendship have been instrumental in carrying out this research from inception to submission.

Many thanks to: Dr. Jenny Freeman who developed the staff training programme and collected data; Dr. Dawn Langdon who performed the neuropsychological assessments; the staff who collected data at the Neurorehabilitation Unit (NRU) of the National Hospital for Neurology in London, the Regional Neurorehabilitation Unit (RNRU) at the Homerton

Hospital in London, and the Rehabilitation Research Unit (RRU) at the University of Southampton; the study coordinators, Ms. Susan Devereux (RNRU) and Mrs. Denise Long (RRU); Dr. Richard Greenwood (RNRU) and Professor Lindsay McLellan (RRU) who enabled this to be a multi-centre study and were Co-Investigators on one of the grants that funded this study: and, importantly, to the patients at the three clinical sites who freely participated in this study, and whose disabilities are sobering for those who treat them.

I am very grateful to: Professor Nick Black who, as part of my advisory committee, provided invaluable advice throughout the study; David Werring, Leslie Duff, Sara Schroter, and Stefan Cano for being extremely helpful; and *Bar Centrale* espresso for helping to generate original thought.

I am indebted to Mum, Dad, Hobey, and Caroline for their support and encouragement which is constant and unconditional.

Finally, my biggest thanks go to Claire, partner, friend, advisor, reviewer, dictionary, thesaurus.....and a whole lot more.

Abstract

Objectives: To evaluate comprehensively the psychometric properties of the Functional Independence Measure (FIM) and the Functional Independence Measure + Functional Assessment Measure (FIM+FAM), and to compare their performance in stroke and multiple sclerosis (MS) patients and with the Barthel Index. To evaluate the conceptual models of both instruments using item analysis, and determine the feasibility of developing a short-form measure. To compare five methods of evaluating responsiveness.

Design: Psychometric study.

Subjects: 209 inpatients with a variety of neurological disorders recruited from three neurorehabilitation units in Southeast England.

Method: Standard methods were used to evaluate the acceptability, reliability, validity, and responsiveness of the FIM and FIM+FAM. Detailed item analyses were performed including internal consistency, intercorrelations between scales and subscales, item convergent and

discriminant validity, and principal components analysis. Item reduction techniques were used to develop a short-form FIM. Five methods were used to evaluate responsiveness: *t* - statistics, relative efficiency, effect size, standardised response mean, and the responsiveness index.

Results: The FIM and FIM+FAM are acceptable, reliable, valid, and responsive measures of disability in neurorehabilitation. However, they demonstrate no psychometric advantage over the Barthel Index, show item redundancy, limited item discriminant validity, and inadequate support for hypothesised subscales. An 8-item short-form FIM is developed that shows similar psychometric performance to the 18-item FIM and 30-item FIM+FAM. Five methods of evaluating responsiveness rank order scales similarly, but generate numerical estimates of different magnitude.

Conclusions: Results demonstrate the need for a more systematic and rigorous approach to the development and psychometric evaluation of instruments before their introduction into practice to ensure the accurate measurement of patient-oriented outcomes in health care. This approach includes the development of appropriate conceptual and measurement models, the application of standard item analysis and item reduction techniques during questionnaire development, and comprehensive evaluation of the recommended full range of psychometric properties.

Table of Contents

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 The public health impact of neurological disease

The public health impact of neurological disease is considerable. It is the single most important group of severely disabling disorders and accounts for over 90% of residents in NHS young disabled units (1, 2). In a health district with a population of 250,000 it has been estimated that between 4,000 and 5,000 people will have a disabling neurological disease (3). Of these, approximately 1,500 people will be so disabled that they require daily help to remain out of institutional care.

The incidence of disabling neurological disorders varies across conditions. At least 130, 000 people in the UK have a stroke each year; of the 80% who survive, many remain permanently disabled (4). Whilst the incidence of other neurological disorders is low, the prevalence is much higher (Table 1.1). This is because many neurological diseases begin in young age and are incurable and progressive over many decades, but have little effect on longevity. They can be complex disorders with diverse effects, an

unpredictable course, and variable manifestations, thus posing unique problems to patients and their families.

Neurological disorders are associated with high health service costs. In the UK, multiple sclerosis (MS) alone is estimated to cost £1.2 billion per year (5), whilst in Sweden (population 9 million), MS is reported to account for 140,000 days absent from work due to sickness per year and 5,048 lost working years due to premature retirement (6). It is notable that the costs associated with chronic disorders increase as disability progresses (7, 8). At present, stroke care accounts for about five per cent of all NHS resources, but this estimate is certain to rise given the increasing incidence of stroke and population ageing (9). As neurological diseases are a major financial concern to the NHS in the UK, rigorous evaluation of the outcomes of therapeutic interventions such as rehabilitation is essential.

1.2 Measuring health outcomes in neurology

Measurement is defined as the assignment of numbers to objects or events according to rules (10, 11). It is an essential component of research in the natural, social, or health sciences (12, 13), and is considered a *sine qua non* of any science (14). In fact, Helmholtz said that "all science is measurement" (cited in 15, page 6).

Measurement in medicine combines laboratory and clinical science. It is aimed largely at determining whether a disease is present (making a diagnosis), and evaluating and quantifying the results (outcomes) of disease and therapeutic interventions. In the laboratory sciences, measurement is based on the development of appropriate instrumentation. Subjective judgement plays a minor role and measurement difficulties are largely amenable to technological solutions. In contrast, clinical measurement relies on human judgement with all its inherent subjectivity (13).

Traditionally, health outcomes have been measured using pathophysiological parameters of disease such as blood tests and radiology, and simple clinical endpoints such as mortality rates and duration of survival (16). Measurement of these indicators requires little subjective judgement. As Streiner and Norman (13, page 1) observe: " 'objective' criteria, based on laboratory or tissue diagnosis where possible, can be used to decide whether a patient has the disease, and warrants inclusion in the study. The investigator then waits an appropriate period of time and counts those who did or did not survive".

Recently, the measurement of health outcomes has become more complex. There is an increasing interest in the measurement of broader, more

abstract, and more subjective aspects of health such as disability, handicap, emotional well-being, health-related quality of life, and patient satisfaction. The interest in measuring broader health outcomes indicates an evolving conceptualisation of health that can be attributed to several factors including: the World Health Organisation (WHO) definition of health as a "state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (17); developments in health care and changing social conditions that have altered disease epidemiology and led to an increase in the prevalence of chronic illnesses (18); the development of new interventions with marginal differences in effectiveness (19); increased demand from commissioners and providers of health care for rigorous evidence of treatment effectiveness (20); and finally, but most importantly, the requirement to incorporate the patient's perspective into health care evaluation (21). The scientific discipline of health measurement has developed in response to the need to supplement clinical judgement about patient outcomes with reliable and valid quantitative measures of aspects of health that were previously thought to be unmeasurable.

A simple but useful classification (see Table 1.2) considers health outcomes in neurology to be either physician or patient-oriented (22). Each of these categories has two subcategories. This classification is not exhaustive but provides a framework for considering health outcomes. Pathophysiological parameters of disease and clinical end-points are termed physician-oriented outcomes because they are defined and measured by clinicians to whom the

results are of most interest. The data generated by these measures of outcome provide information on the presence, natural history, severity, and activity of disease. This information is essential for an improved understanding of disease process and for the evaluation of interventions, and has been the traditional focus of healthcare evaluation.

Physician-oriented outcomes are, however, of limited use when evaluating treatment effectiveness. They do not provide a complete picture of disease impact because they offer limited information about the diverse clinical consequences of disease, and fail to address the effects of illness upon subjectively assessed functioning and well-being (23). There is some evidence that physician-oriented outcomes are not always related to patient-oriented outcomes. Pertinent examples in neurology include the weak relationship between: lesion load quantified by Magnetic Resonance Imaging (MRI) of the brain and disability in MS (24), seizure frequency and aspects of well-being in epilepsy (25), and tremor severity and physical and mental health in Parkinson's disease (26). It is not surprising, therefore, that treatments with demonstrated effectiveness in terms of physiological parameters and clinical endpoints are not always associated with a positive impact on health status. For example, although there is incontrovertible evidence that the use of interferon beta in MS reduces abnormalities detected by MRI and relapse rate (27-29), the effect on disability is unclear and is associated with considerable controversy (30-33).

The limitations of physician-oriented outcomes can be addressed by supplementing these with measures of patient-oriented outcomes. Patient-oriented outcomes are the consequences of disease and treatment that are considered important to patients. Although there is no consensus as to how patient-oriented outcomes should be classified, Gill (34) defines two categories: measures of health status and health-related quality of life. Whilst these two terms have been used interchangeably (35), it is preferable to try and distinguish between them because patient-oriented outcomes are diverse in terms of their methods of administration, the health constructs they measure, and the method by which they are development. In the classification shown in Table 1.2 the term health-related quality of life refers specifically to instruments that are self-report, disease-specific, multidimensional, and incorporate patients in their development. Consequently, it can be misleading for clinicians if all measures that define health beyond traditional indicators of biological function are referred to as measures of health-related quality of life.

Patient-oriented outcomes can be administered using a variety of methods including self-report (also termed patient-based outcomes; 36) and rating by observation (e.g. health professionals or significant others) or interview. Whilst a number of self-report measures exist in neurology, for example the Parkinson's Disease Questionnaire (PDQ-39; 26) and the Subjective Handicap of Epilepsy Scale (SHE; 37), studies have more commonly used

clinician-report measures. For example, studies in MS have mostly used the Kurtzke Expanded Disability Status Scale (EDSS; 38), studies in stroke have used the Rankin Scale (39), and studies in Parkinson's disease have often used the Unified Parkinson's Disease Rating Scale (UPDRS; 40).

Different methods of administering instruments offer unique perspectives and can give conflicting information. This has been demonstrated in a variety of neurological disorders including MS (41, 42), Parkinson's disease (43, 44), epilepsy (45), and stroke (46). As expected, agreement is generally better for directly observable dimensions of health such as physical function than for more subjective dimensions such as pain and emotional well-being (41, 47). Whilst information about outcomes that is obtained from clinician-report is extremely valuable in clinical studies, it does not provide a complete picture of the impact of the disease and treatment on patients as it fails to address subjectively assessed function and well-being (23). As it is only patients themselves who can truly understand the impact of their illness, it is important to incorporate their self-evaluations into any formal assessment of outcomes. Indeed, ignoring the patients' view limits our understanding of the consequences of disease and treatment and limits the ability to help them (48). Consequently, patients' views have been termed the optimum outcome measure (49) and persuasive arguments have been made that all treatment evaluations should include direct reports from patients (50).

Health-related quality of life has been defined in several ways (see 51, page 6, box 2). Whilst there is no agreed definition and considerable controversy as to what the term means (52-55), there is consensus that health-related quality of life is a multidimensional and self-perceived concept. Furthermore, as Fitzpatrick *et al.* (51) observe, the dimensions of a health-related quality of life measure will be disease-specific. For example, patients with Parkinson's disease highlighted the importance of stigma and embarrassment associated with the illness (26). This dimension is not included in generic health measures such as the Medical Outcomes Study 36-Item Short Form Health Survey (SF-36; 56), Nottingham Health Profile (NHP; 57), and EuroQol (58). If these arguments are followed it seems appropriate to reserve the term health-related quality of life specifically for measures that are not only self-report and multidimensional but also disease-specific where the domains are derived from patients with the disorder. Patient-oriented outcomes that do not fulfil these criteria are considered measures of aspects of health status.

In neurology, it is particularly pertinent to measure patient-oriented outcomes as a substantial proportion of neurological disorders are associated with disablement, and many are chronic, progressive, and associated with little prospect of cure. Moreover, advances in basic neuroscience have resulted in the development of therapeutic interventions that either modify disease progression (e.g. interferon beta for MS (59),

antiglutamates for motor neurone disease (60), acetylcholinesterase inhibitors for Alzheimer's disease) or reduce the extent of tissue damage (e.g. thrombolysis in cerebral infarction) rather than arresting the disease process.

In keeping with these developments in neuroscience, neurologists have recognised the importance of measuring patient-oriented outcomes when evaluating therapeutic effectiveness (48, 61). However, there is still strong evidence that physician-oriented outcomes are considered to be more important. In MS , for example, the effectiveness of interferons has been evaluated primarily by measuring disease activity in terms of brain MRI and relapse rate (27, 28) . In stroke, the effectiveness of stroke units has been measured largely in terms of reduction in mortality rates (62), whilst the effectiveness of aspirin has been measured in terms of the prevention of recurrent stroke (63). For epilepsy, seizure frequency has been a prominent outcome in the evaluation of treatment outcomes (25), and in motor neurone disease (MND), the primary outcome in studies evaluating the effectiveness of the antiglutamate riluzole™ was duration of survival (60). Whilst many of these studies have used patient-oriented measures as secondary outcomes, it is notable that few have used patient-report questionnaires and several have concentrated on physical function rather than other domains of health.

The discussion above suggests that neurologists have only recently become interested in patient-oriented outcomes. However, there is considerable evidence to the contrary. For example, in the 1950's Kurtzke (64) developed two disability scales for MS, Rankin (39) developed a disability scale for stroke, and Mahoney and Barthel (65) developed a measure of self-care for patients with neuromuscular and musculoskeletal disorders undergoing rehabilitation. Similarly in the 1960's, Ashworth (66) developed a measure of spasticity for MS, and Hoehn and Yahr (67) developed a severity grading scale for the clinical manifestations of Parkinsonism. The fact that all of these scales are still used widely indicates their clinical usefulness. However, the heavy criticism of such measures (68) indicates that the measurement of patient-oriented outcomes in neurology has progressed slowly over the last 45 years.

Three factors have contributed to this slow progress. First, there have been dramatic advances in basic neuroscience, particularly in neuroimaging, that have maintained the focus on physician-oriented outcomes. For example, the development of computed tomography (CT) scanning in 1972, and soon afterwards MRI, revolutionised the investigation of brain pathology. These techniques, particularly MRI, enable neurologists to quantify the amount of tissue destroyed or disrupted by the disease process. Not surprisingly, research interest has focused on MRI as an objective measure of treatment effectiveness (69) rather than on patient-oriented measures. This situation has been compounded by the absence of gold

standard measures of patient-oriented outcomes (70) and concerns that such data are too soft (71). However, the demonstration of only moderate relationships between pathology and disablement (72), and the demonstration that MRI changes are not accompanied by similar changes in disability, has resulted in a renewed appreciation of the importance of patient-oriented outcomes.

A second reason for the slow development of patient-oriented outcome measures in neurology is that scientific attention has typically focused on the rigour of study design rather than measurement. For example, the MS literature has been more concerned with issues of blinding (73), randomisation (74), placebo-controls (75), sample size (76), ethics (77), and data analysis (78, 79), rather than on the development and evaluation of outcome measures. This bias indicates a failure to appreciate that study design is directly dependent on the properties of the instruments used (80), and that the quality of data is dependent on the quality of the measures used to collect the data (81).

The third and perhaps most important reason for the slow progression of patient-oriented outcome measurement in neurology is due to a general lack of awareness of the science of measurement. Although there is an enormous social sciences literature on methods for rigorously measuring complex constructs, it is not surprising that neurologists are generally

unfamiliar with the fundamental principals of the science of measurement as this literature is outside their domain of expertise. They have, however, appreciated the fundamental importance of rigorous measurement. For example, Kurtzke recognised that disability scales for MS should be user-friendly (64), applicable to all patients (82, 83), and reproducible (64). He stated that the sum total of any patient's disabilities should be able to be categorised on a scale (64), and that any change in disability should be reflected in a change of status on that scale (64). In short, Kurtzke appreciated that scales should be clinically useful (capable of being incorporated into routine clinical practice) and scientifically sound (acceptable, reliable, valid, and responsive). Unfortunately, techniques for evaluating these properties were largely unavailable to clinicians at this time, which is why early scales such as Kurtzke's were not evaluated using standard methods.

1.3 The science of measuring health outcomes

If complex, abstract, and inherently subjective constructs are to be evaluated, then patient-oriented outcomes must be measured. As measurement of such outcomes will influence decisions that affect patient welfare, policy development, and the expenditure of public funds, it is essential that rigorous measurement instruments are used in health care evaluation (84). Because the field of health measurement provides the

scientific basis for evaluating health outcomes, this section provides an overview of this discipline to measuring outcomes in neurology.

Although health measurement as a distinct discipline emerged in the 1980's (97, 111, 113), it is derived from well-established theories and methods of measurement in the field of social sciences whose origins can be traced to the mid 1800's. The basic scientific principles of measurement were established by mathematical psychologists interested in the human being as a measuring instrument. By studying how people make subjective judgements about measurable physical stimuli (e.g. length, weight, loudness), they developed the science of psychophysics: the precise and quantitative study of how human judgements are made (85). The investigation of overt responses to physical stimuli requires precise methods, referred to as psychophysical methods, for presenting the stimuli and for measuring responses (86).

The work of psychophysicists seems far removed from health measurement. In fact, it established the fundamental principles of subjective measurement which are as equally relevant to judgements about health as to judgements about physical stimuli. The psychophysicists demonstrated three important findings about human judgement: that subjective judgement is a valid approach to measurement; that humans make judgements about abstract comparisons in an internally consistent manner; and, that accurate

judgements can be made on ratio rather than simple ordinal scales. It is notable that psychophysical methods are still used in neurology; thermal threshold testing is based on the principle of the just noticeable differences in temperature detection, and audiometry on a person's response to different sound frequencies.

Whilst the psychophysicists were measuring subjective judgements about physical stimuli that could be independently and objectively measured and verified, experimental psychologists were attempting to measure human attributes for which there were no independent physical scales of measurement (e.g. intelligence, personality, attitudes) (86). Darwin's empirical demonstration of evolution in the Origin of Species in 1859 was the impetus behind the study of individual differences in psychology (87). It was reasoned that if animals inherit ancestral characteristics, and if individual differences influence their ability to adapt and survive, so individual differences in humans would have functional significance and could be inherited. Galton, who followed Darwin and believed that the human race could be bettered through controlled mating (eugenics), realised that human characteristics must be measured in a standardised manner before their inheritance could be studied. He coined the term "mental test" for any measure of a human attribute, and set about the large-scale testing of sensory discrimination and motor function in the belief that people with the most acute senses would be the most gifted and most knowledgeable (87). However, when Galton's colleague Pearson

developed and applied the correlation coefficient, it became clear that results from these simple sensory and motor tests bore almost no relationship to measures of intellectual achievement, such as school grades (88). This finding prompted the development of the mental test movement, i.e. the widespread interest in the development and application of mental testing, and the measurement of individual differences.

A major advance in mental testing (12) was made when Thurstone demonstrated that psychophysical scaling methods could be used to accurately measure psychological attributes (89, 90). This finding prompted the development of psychological (or psychometric) scaling methods, which are defined as procedures for constructing scales for the measurement of psychological attributes (85). Spurred on by the practical need to measure diverse outcomes, the mental test movement flourished between 1930 and 1950 with the spread of standardised testing for assessing educational achievement, measuring attitudes and personality, and selecting and screening personnel. In addition, scientific interest in methods of testing led to the development of psychometrics as a prominent discipline within psychology and established the cornerstones of the scientific evaluation of measuring instruments based on reliability and validity testing (85, 91).

The growth and development of psychometrics required standards for the development and evaluation of measurement instruments. The first of these

was introduced in 1954 by a committee of the American Psychological Association (APA; 92). The following year similar guidelines were prepared by a committee representing the American Educational Research Association (AERA) and the National Council on Measurement used in Education (NCME; 93). Subsequently, standards have been published by the Committee to Develop Standards for Educational and Psychological Testing which represents the APA, AERA, and NCME (94-96) along with a commitment to the continual review of measurement standards in psychology and education (96).

Thus, when health care evaluation needed methods for measuring patient-oriented outcomes, the technology already existed. Since the 1970's, the focus of health care evaluation has moved to the measurement of function (the ability of patients to perform the daily activities of their lives), how patients feel, and their own personal evaluation of their health in general (36). The primary source of this information is standardised surveys (97), for which psychometric techniques of scale construction are highly appropriate (36).

Two studies in the US confirmed the value of psychometric methods in assessing health outcomes. The Health Insurance Experiment (98), a randomised experiment conducted by The RAND Corporation between 1974 and 1981, demonstrated that psychometric methods can be used to

generate reliable and valid measures for assessing changes in health status for both adults and children in the general population. Following on from this, the Medical Outcomes Study (36, 99) demonstrated that psychometric methods of scale construction and data collection were successful for measuring health status in samples of sick and elderly people. This study also demonstrated that psychometrically equivalent short-form measures could be constructed from the original longer-forms (100), thereby reducing respondent and administrative burden and improving measurement efficiency. These two pivotal studies confirmed that psychometric methods, borrowed from the social sciences, generated scientifically sound and clinically useful health measures.

Unfortunately, psychometric methods have been slow to transfer to clinical practice. Spitzer (101, page 469-70) argues that this is because "many clinical investigators have not taken the trouble to learn from several decades of conceptual and methodological developments of the social scientist in instrument development". Certainly, many clinicians do not have the time to learn about instrument development and evaluation. In addition, the literature is not easily accessible to clinical investigators. It is directed primarily towards educationalists and psychologists, focuses on issues and topics of little relevance, and is incomprehensible to most readers (13). However, the problem may be more fundamental as many developers and users of health measures are not aware of the field of psychometrics or its relevance to health.

Regardless of the reasons why psychometric methods have been slow to transfer to clinical research, the consequence is that many instruments have not been adequately evaluated from a scientific point of view. Spitzer comments that attempts at validation (when they are present) are often "cursory, ...shoddily executed, ...unsatisfactory, ...unconventional and even bizarre" (101, page 469-470). In his keynote address to the 1986 Portuguese Conference on Quality of Life Research, Spitzer commented: "it is almost amusing to note the plaintive bleat with which authors bemoan the lack of the gold standard. They then go on smugly to do sloppy calibration work which tends to perpetuate the state of affairs" (101, page 470). Spitzer, highlighting the absence of standards for health measurement instruments, called for published minimum criteria of validity agreed by a prestigious panel of inter-disciplinary methodologists, and introduced the notion of a reference library or a bureau of standards to serve all. Such criteria, he says, "should be met in order for people to be credible when declaring a measure valid...we would then end up with a situation where some scientists could advance calibration and others application in a complimentary and synergistic fashion" (101, page 470-471).

These strong words concerning the limited scientific evaluation of many health measurement instruments were important in alerting the clinical community to a crucial yet under-appreciated science. However, Spitzer's

views were perhaps somewhat over-stated as there were already several previous attempts to define frameworks for instrument evaluation and to tackle specific methodological issues (102-108). Subsequently, the mid 1980's saw a rapid expansion of the health measurement literature (13, 51, 109-127). But, as many of these writings are in disparate sources and much of the work represents the opinions of individuals or groups rather than expert consensus, the literature has come under some criticism (128).

In 1994, the Medical Outcomes Trust (MOT) formed an independent Scientific Advisory Committee (SAC) to establish standards for assessing the scientific properties of health measurement instruments (129). This answered the need for consensus standards in health measurement produced by a prestigious inter-disciplinary panel of methodologists, and the development of a bureau of standards. As most users will not have the experience to independently evaluate the measurement qualities of instruments the MOT saw the need to establish the SAC with the aim of facilitating the universal adoption of outcome measures in health care through the use of high quality measurement instruments with user guidelines. The SAC published instrument review criteria (130) in the form of a concise (4 pages), pithy, unreferenced document that provides explicit guidelines for the methods and standards to be used to evaluate the success of instruments according to eight criteria: conceptual and measurement model; reliability; validity; responsiveness; interpretability; burden; alternative forms; and, cultural and language adaptations. These

criteria are now applied to all new measures submitted to the MOT for inclusion in their library.

Since then, two complementary documents (51, 127) have been published in the UK. McDowell and Jenkinson (127) address aspects of instrument design, testing and technical guides, and publication of the completed instrument. They discuss key issues not covered by the MOT instrument review criteria (e.g. item writing, scaling methods, post development marketing), expand on some areas (e.g. conceptual basis for measurement, reliability and validity testing, interpretation of scores), and highlight some important methodological issues.

Fitzpatrick *et al.* (51) were commissioned by the NHS Health Technology Assessment Programme to describe the diversity and reasons for the diversity of available patient-based outcome measures, and to clarify criteria for selecting patient-based outcome measures for use in clinical trials. Fitzpatrick *et al.'s* report is similar to the MOT's instrument review criteria since it discusses eight instrument selection criteria: appropriateness, reliability, validity, responsiveness, precision, interpretation, acceptability, and feasibility. Although it appears that only four criteria are addressed in both documents, more overlap is present. For example, Fitzpatrick *et al.* consider the respondent and administrative burden of health measures under the heading of feasibility, and translation

and cultural applicability under the heading of acceptability. Although Fitzpatrick *et al.* do not provide explicit guidelines for the evaluation of health measures, their document addresses broader issues (e.g. what are patient-based outcomes?), is longer (74 pages), more explanatory, and well referenced. Despite referring to "methods of assessing health, illness, and benefits of health care interventions from the patient's perspective" (51, page iii), Fitzpatrick *et al.'s* criteria are equally pertinent to health outcome measures that, while patient-oriented, are not patient-report (e.g. observer-rated disability measures).

It should be noted that although the principles and standards developed in psychology and education are highly applicable to health, they may not be wholly appropriate as measurement of health differs from measurement of psychological and educational constructs (127). McDowell and Jenkinson argue that "as health is based on biological processes it includes a factual element and a consistent internal logical structure that is absent in ratings of political opinions or economic preferences" (127, page 238).

One measurement property that is specifically pertinent to health measurement is responsiveness: the ability of an instrument to detect change in the outcome of interest (110). As one of the major goals of health care is to change health status, an assessment of the ability of an instrument to detect changes in health is an important prerequisite for the

use of such measures in research and clinical practice (107). However, unlike reliability and validity testing, there is no clear consensus as to how responsiveness should be measured and multiple methods have been proposed in the health measurement literature (110, 116, 131-137). Arguments concern the conditions under which it is appropriate to measure change and the most appropriate approach to the statistical analysis of change (138). It is important (and reassuring) to note that the same arguments took place in the psychology literature many years ago (139, 140) and have not been resolved.

1.4 Measuring health outcomes in neurorehabilitation

Inpatient neurorehabilitation is a therapeutic intervention that provides a co-ordinated multidisciplinary approach to managing the everyday problems of people with disablement associated with neurological diseases. It is recognised as an integral part of health care in neurology and advocated as an important and effective intervention (141-143).

As there is no consensus as to how inpatient rehabilitation should be delivered, there are considerable variations in clinical practice (144-146). However, there is a common conceptual and practical framework to rehabilitation practice that is based on a model of comprehensive care (147). This model emphasises that rehabilitation extends beyond

symptomatic treatment and aims to achieve the optimal quality of life for people within the limits of their diseases (148). Although rehabilitation practice is based on clinical judgement, and its scientific basis is considered to be weak (149), scientific evidence is accumulating that rehabilitation is indeed an effective therapeutic intervention (4, 150).

Despite variations in practice, there is consensus regarding the aims of rehabilitation (144-146, 151). These are: a comprehensive assessment of physical, psychological, and social needs; promotion of physical, psychological, and social adaptation to disability and handicap; facilitation of independence in daily activities; maximisation of patient and carer satisfaction; empowerment; self management; and the prevention of complications. The key elements of the rehabilitation process include a multidisciplinary team approach, individually-tailored programmes, and patient-centred function-based goal setting (143, 152-154).

The WHO's International Classification of Impairments, Disabilities, and Handicaps (ICIDH; 155) is considered to be the cornerstone for evaluating the outcomes of rehabilitation (156). The ICIDH provides a theory of disablement and the rehabilitation process which has proved to be relevant, easily operationalised, and reasonably comprehensive with respect to the aims of rehabilitation. Each of the three concepts can be defined, modified

by treatment, and measured. Furthermore, as the terminology of the ICIDH is well established, data are universally understood (156).

The ICIDH was developed to provide a framework for conceptualising disablement, the consequences of disease (157). In the 1970's, the WHO attempted to investigate how different countries addressed the problems associated with disablement. Data could not be collated as investigators were using neither common terminology nor an agreed set of concepts (158). Recognising the need for a conceptual and operational framework for describing and measuring the consequences of disease, the WHO developed the ICIDH. The conceptual basis of the ICIDH is that the consequences of disease can be defined in relation to different points in its progression from the cause (aetiology), to the active state (pathology) and the long-term consequences on the function of bodily organs (impairment) in terms of functional change (disability) and socioeconomic limitations (handicap) (156).

The ICIDH classifies disablement into three conceptually distinct categories; impairments, disabilities, and handicaps (157). Impairment refer to any loss or abnormality of psychological, physiological, or anatomical structure or function. It represents disturbance at the organ level and indicates that a pathological state is exteriorised (157), that is, "become visible" (159). Disability refers to any restriction or lack of ability to perform an activity in

the manner or within the range considered normal for a human being. It represents the consequences of impairment on functional performance and activity, and indicates that the illness experience is objectified, that is, "externalised" (159). Handicap refers to the disadvantage to an individual that limits or prevents the fulfilment of a role that is normal (depending on age, sex, social, and cultural factors) for that individual. It represents the consequences of impairment and disability on an individual, and indicates the socialisation of an illness experience (155). Since 1993 the WHO have been developing the ICIDH-2: the International Classification of Impairments, Activities, and Participation (160). The revised version avoids the negative connotations associated with the terms disability and handicap. The neutral terms activities and participation are used and the negative circumstances of these dimensions are activity limitations and participation restrictions (159). It is intended that the ICIDH-2 will be completed, approved, and released during 2000 (160).

The conceptual distinction in the ICIDH between impairment, disability and handicap is crucial for two reasons. First, there are clear discordances in severity across the three dimensions. For example, the concert pianist with a broken finger is mildly disabled but greatly handicapped. Second, impairment, disability, and handicap correspond to the obligations of different components of health care (157). Impairment is the primary concern of medical services, disability of rehabilitation services, and handicap of social services (157). Not surprisingly, disability has been the

main health outcome in the evaluation of rehabilitation. Furthermore, physical functioning is fundamental to health care as the degree of disability is a major determinant of the ability to live and work independently and the need for specific kinds of assistance and support (84, 161, 162). Finally, the effect of neurological diseases on neuromuscular function inevitably causes physical disability and, therefore, functional improvement is an important focus for evaluating the impact of neurorehabilitation (149, 163-166).

There are numerous instruments for measuring disability (68, 84, 114, 122, 125, 167, 168). However, it is generally agreed that the field of disability measurement is poorly developed and in its scientific infancy (169, 170). Potential users are faced with a wide choice of scales without adequate criteria for choosing amongst them, and with no consensus as to how disability should be measured (111, 164, 167). The situation is further complicated by conceptual problems in defining disability, the changing emphasis over time of health care for disabled and chronically sick individuals, and a general lack of psychometric expertise in instrument development and evaluation.

At a conceptual level, problems with disability measurement arise from a lack of consensus about definitions, the range of activities to be evaluated, and the operational criteria used. Whilst many definitions of disability have

been proposed (171-173) and there is agreement that disability refers to disease-related restrictions in activity (174), there is no agreement concerning the situations in which these restrictions occur. Some authors (175) define disability as ability without reference to situational requirements (e.g. basic abilities such as reaching, bending, and dexterity). Some authors (176) use the term "functional limitations" to describe such restrictions, and define disability as restrictions in relation to specific domains of a person's own environment e.g. personal care and domestic activities. Some authors (177, 178) extend the definition of disability to include limitations in performance of socially defined roles and tasks within a sociocultural and physical environment.

These disagreements concerning the basic definition of disability have led to an overlap with other health constructs. Although the WHO attempted to separate disability from impairment and handicap (155, 157, 179, 180), these efforts have only been partially successful (181, 182). Some authors argue that assessing limitations in simple activities measures both impairment and disability (183), whilst assessing limitations in complex activities measures both disability and handicap (84, 98).

The activities to be evaluated by disability measures is also an area of controversy. There are an unlimited number of activities and a wide variety

of domains that can be included in disability measures. However, there is no agreement about either (68, 164).

Historically, the development of disability measures since their advent in the 1930's (184) has accompanied changes in the emphasis of health care for disabled individuals rather than developments in the conceptual understanding of disability (84). Early disability scales, such as the Barthel Index (65) and the Katz Index of ADL (185), reflected information requirements for the institutional care of severely disabled individuals by measuring personal activities of daily living (ADL) such as feeding, dressing, and bathing. As the emphasis in health care moved towards care in the community and ultimately social reintegration, measurement of personal ADL was inadequate and wider indicators of applied functional ability were needed. In response to this changing need, disability scales were developed that included instrumental activities of daily living (IADL) such as outdoor mobility, cooking and shopping. Thus, disability scales came to encompass instruments from those limited to personal ADL to more complex measures of social functioning.

The operational definitions that have been used to measure the degree of restriction for activities is another problem in disability measurement. For example, there is a difference between the potential to perform an activity and actual performance. Some authors (186) believe that capacity-oriented

questions are more meaningful in practical terms. Others (84) suggest that performance-oriented questions provide a more realistic assessment of actual disability. Some authors (187) argue for both types of measure as capacity-oriented questions establish the limits but are poor predictors of performance.

Further problems for disability measurement are due to the fact that clinicians are largely unfamiliar with the rigorous scientific methods used to design and evaluate health measurement tools. In order to ensure rigorous measurement of any health outcome, it is essential that the instruments used are scientifically sound (97). As discussed earlier, the theoretical foundations and methodological concepts of measurement were developed in the social sciences, particularly psychology, but have been slow to transfer to medicine. As these techniques remain largely unavailable to clinicians, most disability measures have not been adequately evaluated in terms of their scientific properties. Scale development has been *ad hoc* with little standardisation amongst users and frequent local modifications without a formal evaluation of the scientific properties of the modified instrument.

The Rankin Scale (39) provides an example of local modification of a scale without scientific evaluation. It was developed as a clinician-rated measure of functional recovery after stroke (39). The original publication documents a single-item measure with five response options: no significant disability,

slight disability, moderate disability, moderately severe disability, severe disability. Van Swieten *et al.* (188) modified the Rankin Scale by adding another grade (no symptoms at all), named it the Modified Rankin Handicap Scale, and demonstrated good inter-rater reproducibility (weighted Kappa = .91). Bamford *et al.* (189) also modified the original version of the Rankin scale and called it the Oxford Handicap Scale. They added a new grade (no symptoms), changed the names of the other grades from "disability" to "handicap" (e.g. "moderate disability" to "moderate handicap") and altered the descriptors for each grade so that they were less ambiguous and more focused on handicap. For example, "grade 1, no significant disability: able to carry out usual activities" was changed to "grade 1 minor symptoms that do not interfere with lifestyle". Inter-rater reliability of the Oxford Handicap Scale was reported (weighted Kappa = .72). Neither modification was based on empirically data, and none of the three measures was subjected to adequate psychometric evaluation.

Amongst the many available disability measures, two are becoming increasingly popular among commissioners and providers of health care who have advocated their widespread use for the evaluation of rehabilitation (190). These instruments are the Functional Independence Measure (FIM; 163) and the Functional Independence Measure + Functional Assessment Measure (FIM+FAM; 191). The FIM is popular because it was developed specifically to bring standardisation to disability measurement in medical rehabilitation, was marketed successfully, and was designed to be superior

to alternatives. The FIM+FAM, an extension of the FIM, is popular because it provides a more thorough assessment of cognitive disabilities and is, therefore, most appropriate for disability measurement in patients with neurological diseases. As the widespread use of the FIM and FIM+FAM has important implications for clinical practice, research, and health policy, it is essential that they meet rigorous criteria for both scientifically sound and clinically useful measurement.

1.5 The Functional Independence Measure (FIM) and Functional Independence Measure + Functional Assessment Measure (FIM+FAM)

In 1983 the US Federal Government attempted to introduce mandatory prospective payments for medical services (Social Security Amendments Public Law 98-21 cited in 165). However, the absence of uniform and reliable methods of measuring the outcomes of chronic care contributed to the exemption of medical rehabilitation facilities from this system (192, 193). Recognising the inability to evaluate and compare different clinical practices in rehabilitation, a Task Force was formed to develop a Uniform National Data System for Medical Rehabilitation (UDS). The purpose of the UDS was to improve the effectiveness and efficiency of rehabilitation services in the United States. The FIM (Table 1.3) was developed as the disability measurement instrument of the UDS (163, 194). In attempting to standardise rehabilitation practices, the FIM has succeeded in becoming

widely used throughout the world, not only as part of the UDS, but also as a stand alone measure of disability. By 1993, FIM data on 150,000 patients had been entered into the UDS database (195).

In developing the FIM, the Task Force reviewed 36 published functional assessment instruments and other unpublished instruments that would be helpful in identifying items that measure function. From this process, the Task Force selected the most common and useful functional assessment items (163). Initially 16 items and a four-point response scale were chosen. This instrument was piloted for face validity and ease of administration (163). Following this process one item was discarded, two items were added, one of the response options was divided into three intermediate levels, another response option was added, and minor (unexplained) revisions were made (163). This version of the FIM was then used in the trial phase to evaluate inter-rater reliability, validity, precision, and time and cost to administer. Subsequently, another item has been added.

The FIM+FAM (Table 1.4) is an extension of the FIM. Clinicians using the FIM in brain injured patients felt that the five items of the FIM cognitive scale provided only a modest assessment of communication, cognitive, and psychosocial disabilities (191). In order to provide a more thorough evaluation of these domains of disability, a group at the Santa Clara Valley Medical Centre developed a 12-item adjunct, the Functional Assessment

Measure (FAM) (191). The FAM items (nine to assess cognitive disabilities and three to assess motor disabilities) were designed to be added to the FIM in order to produce the FIM+FAM, a more comprehensive and sensitive instrument (191). They were not intended to be a stand alone measure. Like the FIM, the FIM+FAM has been received enthusiastically, reflecting the need for a measure of cognitive disabilities. Although it was originally intended as a measure for neurological disabled patients with brain injury, the developers of the FIM+FAM now recommend its use for all patients undergoing rehabilitation (Karyl Hall personal communication July 1997).

1.5.1 Description

The FIM (Table 1.3) is an 18-item instrument that measures disability in terms of burden of care. The items comprise two scales, a motor scale containing 13 items and a cognitive scale containing five items. Each of the two FIM scales has two or more subscales. The motor scale has four subscales: self-care (six items), sphincter care (two items), transfer (three items), and locomotion (two items). The cognitive scale has two subscales: communication (two items) and social cognition (three items). Each item has a seven-point response scale (Table 1.5: 1 = maximum disability, 7 = minimum disability) and is rated by consensus opinion of the rehabilitation team treating the patient on the basis of observing patient behaviour for up to 72 hours. Items are added to generate summary scores for the six subscales, motor and cognitive scales, and a total score. The FIM is

designed to be discipline-free and, therefore, can be used by any trained rehabilitation professional (163, 165, 166).

Guidelines for rating FIM items are contained in the manual (196). For each item there are two aids for rating: a written text and a flow diagram (decision tree). Appendix 1 provides the guidelines for rating the grooming item and illustrates how they are individually tailored.

The FIM+FAM (Table 1.4) is a 30-item instrument with the same conceptual basis and 7-point response scale as the FIM. The items comprise two scales, a motor scale containing 16 items and a cognitive scale containing 14 items. Each of these two FIM+FAM scales has three or more subscales. The motor scale has four subscales: self-care (seven items), sphincter care (two items), transfer (four items), and locomotion (three items). The cognitive scale has three subscales: communication (five items), psychosocial adjustment (four items), and cognitive function (four items). Ratings are made by team consensus from behavioural observation for up to 10 days and items are added to generate summary scores for the seven subscales, motor and cognitive scales, and a total score. Like the FIM, the FIM+FAM is designed to be discipline-free so that any trained rehabilitation professional can administer the scale (Karyl Hall, personal communication 1993). As for the FIM, the common 7-point response scale

is individually tailored for each item. A manual is available to guide rating of the 12 FAM items.

1.5.2 Clinical usefulness

By 1993 the clinical acceptance of the FIM was not in doubt. In the US, it was being used in approximately 60% of rehabilitation facilities and already the Uniform Data System had collected FIM data for nearly 150, 000 patients (195). In addition, two studies have indicated that the FIM can be administered by team consensus (197) or by an individual rater (163) in less that 15 minutes. Less data were available for the use of the FIM+FAM in the US, or the FIM and FIM+FAM in the UK. However, the first FIM and FIM+FAM workshop in the UK in December 1994, organised by the British Society of Rehabilitation Medicine (190), was attended by over 400 delegates. A UK FIM+FAM Users Group was formed in 1995 to develop guidelines for use of both instruments and provide regular training workshops. This Group continues to expand (personal communication 1999: FIM+FAM co-ordinator, Regional Neurorehabilitation Unit, Northwick Park Hospital, Harrow, Middlesex).

1.5.3 Scientific rigour

In the 15 years since the FIM was developed and 10 years since the FIM+FAM was developed, the field of health measurement has advanced considerably. It is understandable, therefore, that the examinations of the measurement properties of the two instruments reported in the original publications (163, 191) now appear somewhat limited. Even though data have been reported subsequently, when these are compared with recommended standards both instruments remain poorly evaluated from the scientific perspective.

The original publications about the FIM and FIM+FAM reported little information about their scientific properties. In fact, for the FIM, no empirical data were reported but future psychometric analysis plans were outlined (163, 165, 166). For the FIM+FAM, evidence for reliability was reported in terms of percent agreement for item scores generated by 20 observers from three patient narratives (191). Results demonstrated considerably worse agreement for the FAM items than for the FIM items. Even though this statistical method does not correct for chance agreement (198, 199), it raised concerns that the inter-rater reliability of the FIM+FAM might be less than acceptable. Evidence for the validity of the FIM+FAM is also found in Hall *et al.'s* study (191). Although not stated explicitly by the authors, correlations between the FIM+FAM and other measures provide evidence of convergent and discriminant construct validity.

Subsequent to their use in clinical practice, data supporting the scientific properties of the FIM and FIM+FAM have accumulated. For the FIM, evidence supports its reliability (194, 195, 200-207), validity (191, 197, 201, 207-212) and, latterly, its responsiveness (207, 213). For the FIM+FAM, although studies have addressed the reliability of the items (214, 215) and validity of the total score (216), little is known about its psychometric properties. The results of these studies are discussed in more detail later.

Despite the accumulation of evidence for the psychometric properties of the FIM and FIM+FAM, the evaluation of both instruments is limited when these studies are compared with the standards recommended by the Medical Outcomes Trust (130), McDowell and Jenkinson (127), and Fitzpatrick *et al.* (51). Although it is widely known that measurement properties are largely independent of each other (107, 217), but dependent on the sample in which they are examined (51, 127, 130), only one recent study of the FIM has examined some aspects of its reliability, validity, and responsiveness in the same sample (207). In fact, there have been no comprehensive psychometric evaluations of either the FIM or FIM+FAM. Similarly, despite multiple reliability studies of the FIM, no study has comprehensively evaluated all types of reliability that are appropriate for multi-item observer-rated measures. Likewise, there are no comprehensive validity studies for either measure. Those reported have concentrated on convergent validity

rather than proposed a validation strategy based on explicit logic (127), and examined the extent to which empirical data supported hypotheses concerning the behaviour of the measure and its components. There are few responsiveness data.

Further studies of the FIM and FIM+FAM are required. For example, none have compared the performance of the FIM or FIM+FAM in different disease groups. Although designed to be generic measures, and theoretically usable with any disorder in many different settings (218), this assumption has not been tested for either the FIM or FIM+FAM. This is particularly important as both instruments cover a limited range of the disability spectrum. Similarly, no studies have examined whether the FIM is superior to the Barthel Index developed more than 30 years previously, or whether the FIM+FAM is superior to the FIM in neurologically disabled patients. To justify their introduction into clinical practice, new instruments need to demonstrate superior measurement properties to existing measures. Also, no studies have examined the extent to which empirical evidence supports the conceptual models of disability defined by the FIM and FIM+FAM. Examination of the scale and subscale structure of a measure, and the procedures followed to create scale and subscale scores, is necessary to justify the selection and grouping of items and the reporting of summary scores (130). This is important given concerns raised by others about the conceptualisation of disability (111, 164). Finally, the feasibility of developing short-form versions of the FIM or FIM+FAM has not been

examined. As patients are sick or disabled, and staff resources are limited, it is necessary to maximise the measurement efficiency of an instrument: maximum information from the minimum number of items (219).

1.6 Summary and study objectives

Disability due to neurological disease has a considerable public health impact which justifies the importance of measuring disability as an outcome of neurorehabilitation. Rigorous disability measurement can be achieved using psychometric methods of scale construction. The FIM and FIM+FAM have an important role in the future of disability measurement and the evaluation of rehabilitation. However, their psychometric properties have not been extensively studied. Responsiveness is an important property of health measures. Unlike reliability and validity, there is no consensus as to how it should be measured.

The first objective of this study is to evaluate comprehensively the psychometric properties of the FIM and FIM+FAM. This includes a comparison of the performance of both instruments in stroke and MS patients and with the Barthel Index. The second objective of this study is to evaluate conceptual models of disability through detailed item analyses of both measures. This includes an examination of the feasibility of developing

a short-form measure. The third objective of this study is to compare methods of evaluating responsiveness.

Chapter 2

General Methods

2.1 Study sites

Patients were recruited to the study at three clinical sites: the Neurorehabilitation Unit (NRU), National Hospital for Neurology and Neurosurgery; the Regional Neurorehabilitation Unit (RNRU), Homerton Hospital; and the Rehabilitation Research Unit (RRU), Southampton General Hospital.

These three neurorehabilitation units treat different patient populations. The NRU treats predominantly patients with multiple sclerosis (40% of admissions in 1993) whereas the RNRU treats predominantly patients with head injury, and the RRU treats predominantly patients with complex stroke. At the NRU and RNRU the second most common diagnostic group is complicated stroke; at the RRU this is head injury. All three units provide inpatient rehabilitation for any neurological disorder in all age groups and all are allied to regional neurosurgical units. Collaboration of these three units aimed to ensure that this study encompassed a wide spectrum of neurological disease, addressed high incidence pathologies and common

causes of severe and progressive disability in predominantly young people, maximised patient numbers, and included a broad range of rehabilitation programmes.

2.2 Recruitment

The recruitment strategy differed at each unit due to the varying rates of patient turnover. Methods of recruitment were pre-determined to limit selection bias. At the NRU, a maximum of two patients was entered into the study every Monday over a period of 18 months. The first two patients who arrived on that day were selected. At the RNRU, a maximum of one patient each week was entered into the study until the target number of 60 subjects was attained. Of the planned admissions each week, the person whose surname was nearest to the beginning of the alphabet was selected. At the RRU, all patients admitted over a one-year period were invited to participate in the study.

Ethical approval was obtained from the ethical committees of each study site and informed consent was obtained before any patient was enrolled. In circumstances where patients were not able to give informed consent (e.g. due to cognitive impairment or aphasia), written consent was obtained from the next of kin. Any patient over 16 years of age, with any neurological disorder, who consented to participate was eligible for entry into the study.

Patients were excluded if they declined to participate, were admitted for respite care rather than rehabilitation, or had an expected duration of stay of less than two weeks. Appendix 2 contains the ethical approval, consent form, and patient information leaflet for each of the three clinical sites.

2.3 Rehabilitation intervention

Each of the three units provides intensive, multidisciplinary, goal-oriented, inpatient rehabilitation. Whilst standard techniques and methods are used, the nature of the rehabilitation process is tailored to individual patients according to diagnosis, disabilities, handicaps, needs, and goals. Rehabilitation might include any combination of the following professional disciplines: medicine, physiotherapy, occupational therapy, neuropsychology, speech and language therapy, nursing, and social work.

On admission to the rehabilitation unit, each patient is assigned to a treating team consisting of a member from each professional discipline required for the rehabilitation treatment plan. A typical team consists of a nurse, physiotherapist and occupational therapist. Neuropsychologists, speech and language therapists, and social workers are part of the treating team when appropriate. Treating teams range in size from three to six persons and are responsible for consensus rating of the FIM, FIM+FAM, and Barthel Index.

2.4 Health outcome measures

Table 2.1 provides details about the outcome assessment in this study, including the method of administration, assessment point, and site of administration for all health outcome measures.

2.4.1 FIM and FIM+FAM

The FIM (163) and the FIM+FAM (191) measure disability in terms of independence in functional tasks. Both measures are fully described in Chapter 1.

2.4.2 Barthel Index

The Barthel Index (65, 220) measures disability as independence in 10 personal activities of daily living such as feeding, dressing, and bathing (Appendix 3). Items are rated on a 2-point (two items), 3-point (six items), or 4-point scale (two items) and summed to give a total score ranging from 0 (maximum disability) to 20 (minimum disability). Rating is from observation by any health professional. In this study the Barthel Index is rated by team consensus opinion of the treating multidisciplinary team of each patient.

The Barthel Index is widely used as a measure of disability, has been described as "the best activities of daily living measurement scale" (221), and is recommended as a benchmark against which other instruments should be evaluated (149). Multiple versions of the original instrument exist (222-225). Wade's version (226) is used in this study as this is the one advocated by the Royal College of Physicians of London (227) and the version on which most psychometric data are available.

A number of studies have addressed the reliability (221, 223, 224, 228-232), validity (220, 225, 233-235), and responsiveness (236) of the Barthel Index. Although the Barthel Index is widely regarded as a reliable and valid measure of disability, a closer analysis of the data shows that the psychometric properties have not been comprehensively evaluated and different studies apply to different versions of the instrument. Notable deficiencies are in the assessment of construct validity and responsiveness which have received little attention. However, the available data are very encouraging.

2.4.3 Modified Barthel Index

The Modified Barthel Index (237), shown in Appendix 4, is an alternative form of the Barthel Index. This was used at one site (RNRU) where it was

developed to meet the needs of their patient population and is in routine clinical practice. The only modification involves changes in the guidelines for scoring; the 10 items and rating scale of the Barthel Index remain unchanged. Preliminary psychometric data suggest that the modified Barthel Index retains validity and inter-rater reliability (237). Modified Barthel Index ratings were made at the same time and in the same manner (consensus opinion of the treating team) as FIM and FIM+FAM ratings.

2.4.4 Kurtzke Expanded Disability Status Scale (EDSS)

The EDSS (38) is an MS-specific, neurologist rated instrument grading disability on a continuum of 0 (normal neurological examination) to 10 (death due to MS) in 20 steps (Appendix 5). It is scored on the basis of the neurological history and examination and was developed specifically to enable comparisons of disability within and between patients. The EDSS is the most widely used measure of outcome in clinical trials of MS (27, 28, 238). Kurtzke also developed the Functional Systems (FS, 82) which consists of eight scales representing different functions of the central nervous system: pyramidal, cerebellar, brainstem, bladder / bowels, sensory, mental, visual, and other. The FS and EDSS are intimately related. The FS delineates the type and severity of eight neurological impairments and the EDSS represents the sum of a person's neurological dysfunction (82). Hence, comments about FS scores appear in the guidelines for rating the EDSS.

Studies evaluating the psychometric properties of the EDSS report variable reliability (207, 239, 240), support convergent validity (207), and demonstrate limited responsiveness (70, 207). However, a close examination of the literature indicates that no comprehensive evaluations of the EDSS have been undertaken.

In this study, EDSS and FS data were collected only at the NRU as this was the only unit regularly treating patients with MS. All ratings were undertaken by a single neurologist (JH) on the basis of clinical examination and patient interview. Only EDSS data are reported.

2.4.5 Office of Population Censuses and Surveys Disability Scales (OPCS)

The OPCS (175) measure disability in 13 dimensions: locomotion, reaching and stretching, dexterity, personal care, continence, seeing, hearing, communication, behaviour, eating / drinking / digestion, disfigurement, intellectual functioning, and consciousness (Appendix 6). Disability in each dimension is graded on an individually weighted scale and is rated from patient interview. OPCS scores can be reported in three ways: 13 scores for the individual dimensions of disability, an overall weighted disability

severity score, and an overall disability severity category. In all circumstances high scores indicate high disability. The weighted overall disability severity score ranges from 0.5 to 21.4 and is computed from the three highest ratings on the 10 core dimensions (eating / drinking / digestion, disfigurement, and consciousness dimensions are excluded). This severity score then translates to a disability severity category between 1 and 10 in single point increments.

OPCS scales were developed for use in a national UK survey to investigate the prevalence and severity of all forms of disability in the adult population (175). The scales are based on the conceptual framework of the WHO International Classification of Impairments, Disabilities, and Handicaps which categorises (155).

A limited psychometric evaluation of the OPCS scales has been undertaken (234). Inter-rater reliability between two independent raters was high (r = 0.96; n = 120). The type of correlation coefficient used is not reported. Evidence for convergent construct validity was provided by demonstrating a high correlation with the Barthel Index (*rho* = 0.82; n = 265). Evidence for comparable responsiveness of the OPCS and Barthel Index was provided by demonstrating significant improvements in mean scores for two groups of patients whose level of disability was expected to change, but not in a third group whose level of disability was not anticipated to change.

In this study OPCS disability data were collected on a subsample of patients involved in the study at two units, NRU and RRU. Ratings were based on patient interview with the study co-ordinator at the relevant unit. Overall weighted disability severity scores are reported.

2.4.6 London Handicap Scale (LHS)

The LHS (241) is a self-report generic measure evaluating handicap as degree of disadvantage on six items: mobility, physical independence, occupation, social integration, orientation and economic self-sufficiency (Appendix 7). Each item is rated on a 6-point scale (1 = minimum handicap, 6 = maximum handicap) and raw scores are weighted using part utilities. Item scores can be reported as a profile of disadvantages or summed to generate an overall handicap severity score.

The LHS shows adequate internal consistency (alpha coefficients: .67 to .88) and test-retest reproducibility (Intraclass correlation coefficients: .72 to .91) for group comparison studies. Evidence supports content validity, construct validity (internal consistency, factor analysis, group differences, convergent and discriminant validity). Responsiveness was determined by

examining pre and post intervention scores in nine studies, effect sizes ranged from 0.07 to 0.85 (241).

The manual provides extensive information on the development and evaluation of the instrument as well as comprehensive guidelines for its use and the interpretation of data (241). The LHS has been approved by the Medical Outcome Trust (242).

2.4.7 Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36)

The SF-36 (56) is a self-report, generic measure of health status in eight dimensions (Appendix 8). Two summary measures, the Mental and Physical Component Summary Scales, can also be generated (243). The reliability and validity of the eight dimensions and two summary measures have been extensively evaluated with favourable results and are summarised elsewhere (56, 243-247).

Scores for the eight SF-36 scales range from 0 (poorest health) to 100 (best health) (56). The two summary measures are scored to have a mean of 50 and standard deviation of 10 in the general US population (248). SF-36

data were only collected at the NRU and RRU. In this study summary scores are reported used.

2.4.8 General Health Questionnaire (GHQ)

The GHQ (249) is a self-report measure of psychological distress (Appendix 9). The 28-item version used in this study has four subscales each containing seven items: somatic symptoms, anxiety and insomnia, social dysfunction, and severe depression. Items are rated on a dichotomous rating scale and summed to generate subscale and total scores (250). High scores indicate greater psychological distress. Evidence supports the reliability, convergent and discriminant validity of the GHQ-28. Several other studies address the validity of the various versions of the GHQ (84, 114, 249, 250). No responsiveness data are available.

2.4.9 Measures of neuropsychological functioning

Measures of neuropsychological functioning included two measures of global cognitive decline, three measures of reasoning ability, and two memory measures. The two measures of global cognitive decline were:

2.4.9.1 Mini-Mental State Examination (MMSE)

The MMSE (251) measures cognitive state on the basis of 11 items in five domains: orientation (2 items), registration (1 item), attention and calculation (1 item), recall (1 item), and language (6 items) (Appendix 10). The MMSE, designed as a simplified form of the cognitive mental status examination, concentrates only on the cognitive aspects of mental functions and is in widespread clinical use. The MMSE is rated by interview, with the points scored for correct responses summed to generate a total score ranging form 0 (maximum cognitive impairment) to 30 (no cognitive impairment registered on the scale). Values less than 24 are considered indicative of cognitive impairment.

Limited psychometric data are available for reliability and validity. High levels of test-retest reliability (r = .89 to .99) and inter-rater reliability (r =.83) have been reported (251). High correlations between the MMSE and the verbal IQ (r =.78) and performance IQ (r =.67) of the Wechsler Adult Intelligence Scale provides evidence for convergent construct validity. Evidence supports group differences construct validity of the MMSE. Responsiveness has been evaluated by examining change scores pre and post treatment. For patients with uncorrectable brain disease (dementia), there was no significant change between pre and post treatment results. A small but significant difference is shown in patients with depression, and a

large and significant difference is shown in patients with depression associated with severe cognitive impairment (251).

2.4.9.1 Wechsler Adult Intelligence Scale-Revised Version (WAIS-R)

The WAIS-R (252) is a measure of general intellectual level. It comprises 11 subsets divided into verbal (six subsets) and performance (five subsets) subscales. Scores are derived for verbal, performance, and full scale IQ. The WAIS-R gives a good indication of the current level of intellectual function (253), with high scores indicating high intellectual performance. Only the verbal subscale (information, comprehension, vocabulary, similarities of pairs of words, arithmetic, and digit span) was used as neurological disability can affects patients on the performance subtest (253). There is good support for the reliability and validity of the WAIS-R (252).

The three measures of reasoning ability were:

2.4.9.3 Halstead Book Category Test (HBCT)

The HBCT (254) is a measure of abstracting ability (reasoning) and consists of 208 visually presented items in seven subsets. For six of the subsets, items are organised on the basis of different principles. The subject's task is to figure out the principle presented in each set and signal the answer. The

seventh subset is made up of previously shown items and tests the subject's recall. The score is the number of errors. Therefore, high scores indicate poor reasoning ability. A number of studies demonstrate the reliability and validity of the HBCT (254).

2.4.9.4 Wisconsin Card Sorting Test (WCST)

The WCST (255) is a measure of reasoning, specifically abstract behaviour and set-shifting ability. Subjects are given a pack of 64 cards on which are printed one to four symbols (triangle, star, cross, circle) in one of four colours (red, green, yellow, blue). No two cards are identical. The subject's task is to place the cards, one by one, under four stimulus cards (one red triangle, two green circles, three yellow squares, and four blue stars) according to the principle that the subject must deduce from the examiner's responses. The test begins with colour, then shifts to form, then to number. High scores indicate good reasoning ability. The reliability and validity of the WCST have been demonstrated in a number of studies (255).

2.4.9.5 Verbal and Spatial Reasoning Test (VESPAR)

The VESPAR (256) is a test of reasoning specifically designed for patients with neurological disability. Three types of inductive reasoning are

examined: categorisation, analogy, and series completion (253). Each of these three problems is arranged in matched sets of 25 verbal and spatial items. The matched design allows clear conclusions to be drawn if either verbal or spatial stimuli lead to poor performance. This is because the difference is unlikely to be due to different test procedures or task demands and most likely to be due to a specific deficit in either verbal or spatial reasoning (253). The stimuli were selected for their appropriateness for neurological patients. The verbal items use words less vulnerable to acquired language deficits and the spatial items do not depend on fine visual acuity or shape discrimination. No manual dexterity is required and there are no penalties for slow performance (256). High scores indicate good reasoning ability. Evidence supports the reliability and validity of the VESPAR (256).

The two measures of memory were:

2.4.9.6 California Verbal Learning Test (CVLT)

The CVLT (257) is a measure of interaction between verbal memory and conceptual ability. It provides information about a subject's use and effectiveness of learning strategies and the capacity for concept formation (258). Subjects are presented with lists of 16 items and the number of items recalled is counted. High scores indicate good memory function. Several

studies have demonstrated the reliability and validity of the CVLT (257, 258).

2.4.9.7 Recognition Memory Test (RMT)

The RMT (259) measures two aspects of memory: verbal (recognition memory for words) and non-verbal (recognition memory for male faces) memory. Each test contains 50 stimulus items and 50 distractors. Only the recognition memory test for words was used in this study as neurological disorders often result in visual disabilities. First, subjects are presented with the 50 one-syllable high frequency stimulus words (one every three seconds) and required to say whether they like each word as a method of ensuring their attention. Then, 50 pairs of words are shown which include one of the original words. They are required to identify the word seen previously. Recognition Memory Tests are a relatively pure test of memory function as they place few demands on other cognitive functions (253). High scores indicate good memory function. Reliability and validity have been demonstrated for the RMT in a number of studies (258, 259).

All measures of neuropsychological functioning, except for the MMSE, were administered only at the NRU as this was the only clinical site with a full-time neuropsychologist. One neuropsychologist administered all these measures. The MMSE was administered by the study co-ordinator at two

sites (NRU and RRU) on admission and discharge. MMSE scores of less than 18 were used as a cut-off to evaluate whether a patient would be able to adequately complete self-report questionnaires.

2.4.10 Staff-report transition question of change in disability

Staff perceptions of change in disability due to rehabilitation were measured using a transition (131, 260; Appendix 11). On discharge from each neurorehabilitation unit, the treating multidisciplinary team for each patient was asked to indicate on a 7-point scale (1 = marked deterioration, 4 = no change, 7 = marked improvement) the extent of their change in disability due to rehabilitation.

2.5 Training of FIM and FIM+FAM raters

In the few days before the study commenced, FIM and FIM+FAM raters at the three study sites underwent a structured training programme that was developed locally at the NRU. All three clinical sites were already using the FIM routinely. The formal training programme lasted one day. First, there was lecturer on the importance of measuring outcomes, the development of the FIM and FIM+FAM, and the need for this study. Next, there were four training sessions: basic scoring principals, scoring cognitive items, scoring communication items, and scoring physical and self-care items. Each of

these interactive sessions consisted of vignettes (simple, short, written scenarios of patient performance) based on actual clinical examples and video footage of patients whose consent had been obtained in advance. Ongoing training of new staff was undertaken by the study co-ordinator at each clinical site. Appendix 12 reports the results of a small study, undertaken during training, to determine the effect of the training programme on rater proficiency.

2.6 Data collection

Recruitment at NRU took place between June 1994 and February 1996 (21 months); at RNRU between June 1994 and October 1995 (17 months); and at RRU between July 1995 and May 1996 (11 months). At each study site data were collected on admission and discharge. Collection and storage of data were the responsibility of each study co-ordinator. Details about the outcome measures used are provided in Table 2.1.

Within 48 hours of admission to the neurorehabilitation unit, consenting patients were interviewed by the study co-ordinator to obtain sociodemographic and diagnostic data. Where necessary (e.g. diagnosis), these data were substantiated by review of the medical notes. At this interview, instruments rated by the study co-ordinator (e.g. MMSE, EDSS) were administered and self-report questionnaires (e.g. SF-36) were

distributed. Instruments rated by consensus opinion of the treating multidisciplinary team (e.g. FIM and Barthel Index) were rated independently from other admission data at a team meeting. The timing of the team meeting to decide consensus rating differed for each unit. Team ratings were made three or four days after admission at the NRU, two days after admission at the RNRU, and seven to ten days after admission at the RRU. These differences reflected the usual clinical practice of each unit.

Within 72 hours of discharge, the study co-ordinator at each unit distributed self-report questionnaires and administered co-ordinator rated measures. Measures rated by team consensus opinion were collected independently at a discharge meeting along with the transition question. Raters did not have access to admission scores. As the SF-36 concerns health status "in the past four weeks" discharge SF-36 ratings were collected by postal survey four weeks after the date of discharge.

All ratings made by team consensus opinion did not include the study co-ordinator. All information was collected on separate sheets of paper, gathered and stored by the unit co-ordinator. Admission data were not available for review at discharge. Where guidelines are available, all measures were administered in accordance with developers recommendations.

Chapter 3

Psychometric Evaluation of the FIM and FIM+FAM: Method and Results

This chapter reports the method and results for the psychometric evaluation of the FIM and FIM+FAM. This chapter includes: a comprehensive evaluation of the acceptability, reliability, validity, and responsiveness of the FIM and FIM+FAM; a comparison of the performance of the FIM and FIM+FAM in patients with stroke and with MS; and a comparison of the psychometric properties of the FIM, FIM+FAM, and Barthel Index.

3a Method

3a.1 Psychometric evaluation of the FIM and FIM+FAM

Standard psychometric methods are used to determine the acceptability, reliability, validity, and responsiveness of the FIM and FIM+FAM. Identical analyses are used for both instruments and all analyses are performed separately for total, motor, and cognitive scales.

3a.1.1 Acceptability

An instrument is considered acceptable when it can be successfully incorporated into clinical practice and when score distributions adequately represent the true distribution of health status in the sample (261). An empirical indicator of whether an instrument can be incorporated into clinical practice is proportion of missing data which is calculated as the percentage of possible responses that are missing (262). To the extent that the criteria defined below concerning score distributions are satisfied, items and scales are considered acceptable.

Score distributions are reported for FIM and FIM+FAM item and scale ratings at admission. Item score distributions are considered acceptable when four criteria are met: all response categories are endorsed (ideally with equal numbers of patients endorsing each response option for maximum discrimination (85); maximum endorsement frequencies, calculated as the percentage of responses for the most frequently endorsed response category, do not exceed the generally recommended criterion of 80% (13); and item floor and ceiling effects, calculated as the percentage of responses for the lowest and highest scores, respectively, are minimal. Widely accepted criteria for maximum item floor and ceiling effects do not exist. Two published recommendations are 75% (263), and 90% (264). However, the choice of neither is substantiated. As maximum endorsement

frequency and item floor and ceiling effects are logically related, the criterion of 80% was chosen for this study.

Scale score distributions are considered acceptable when: scores span the full scale range (97); mean scores are situated near the scale mid-point (265); scale floor and ceiling effects, calculated as the percentage of responses for the minimum and maximum scores, respectively, are minimal; and score distributions are not excessively skewed (246). There are no widely accepted criteria for maximum floor and ceiling effects and extent of skewness for scales. Authors have recommended that scale floor and ceiling effects should not exceed 15% (266) or 20% (267). In this study the more stringent criterion of 15% is chosen. Holmes *et al.'s* recommendation that skewness statistics should be within the -1 to +1 range (263) is adopted.

3a.1.2 Reliability

In classical test theory it is assumed that the scores generated by a measurement instrument, observed scores, include two components: a true score and random error (217). The relationship between observed scores (o), true scores (t), and random error (e) is expressed as:

$$O = t + e$$

The reliability of an instrument is defined as the extent to which it is free from random error (85). As reliability increases (or decreases), scores are more (or less) consistent and, therefore, measured variance reflects true variance in the construct (or random error) (47). In keeping with this definition, reliability coefficients estimate the proportion of total score variance that is due to true score variance (85).

In this study two types of reliability are examined: internal consistency and reproducibility. These are the most appropriate methods for estimating the reliability of multi-item observer-rated instruments like the FIM and FIM+FAM (268). In addition, standard errors of measurement are calculated from reliability coefficients to determine confidence intervals around individual patient scores.

3a.1.2.1 Internal consistency

Internal consistency is the extent to which items are interrelated (217). If the items in a scale are assumed to measure the same underlying construct, the intercorrelations among the items represent the reciprocal of error (269). Three measures of internal consistency are examined: corrected item-total correlations, Cronbach's alpha coefficients, and homogeneity coefficients.

Corrected item-total correlations are Pearson's product-moment correlations between item and scale scores after the item of interest has been removed to prevent spurious estimates (85, 270). These analyses indicate the strength of relationship between item and scale scores. The higher the correlation, the higher the shared variance and the higher the reliability of the item. A range of recommended minimum values for item-total correlations has been suggested: .20 (113); .30 (271); .40 (246). In this study .30 is used as the minimum criterion for item-total correlations.

Cronbach's alpha coefficients estimate the internal consistency of a group of items from their average intercorrelation (272). Alpha coefficients indicate the proportion of the variance of the sum of the items which is comprised of the sum of the covariances of each pair of items (273). Alpha coefficients exceeding .70 are considered acceptable for scales used to make group comparisons, whereas the more stringent criterion of .90 to .95 is required for scales used to make individual comparisons (130, 217). In this study alpha coefficients are reported with single sided (lower limit) confidence intervals (273, 274) to determine the likelihood that obtained values are significantly greater than the recommended criteria (273). Alpha coefficients are also reported for scales with items omitted one by one to assess the influence of individual items on internal consistency.

As alpha coefficients are related to scale length (275-277) Ware *et al.* (97) recommend that homogeneity coefficients are also reported as indices of internal consistency. Homogeneity coefficients are the average item-intercorrelations for scales; it is recommended that values exceed .30 (265). They are of particular value when comparing the internal consistency of instruments with differing numbers of items.

3a.1.2.2 Reproducibility

Reproducibility is the agreement between two or more ratings of the same patient (130, 271). Two types of reproducibility are examined: intra-rater and inter-rater.

Intra-rater reproducibility is the agreement between two or more ratings made by the same observer of the same patient. It provides an estimate of the stability of scores and within-rater variability over time. For the FIM and FIM+FAM, intra-rater reproducibility is estimated by determining the agreement between ratings made by the same multidisciplinary team for the same patients on two different occasions. The interval between the two ratings is three to five days. This interval was chosen specifically to minimise rater memory bias and the likelihood of change in disability between the two ratings. These influences tend to over- and underestimate instrument reproducibility, respectively (278).

Inter-rater reproducibility is the agreement between ratings on the same patients made by different observers at approximately the same point in time. It provides an estimate of between-rater variability in scoring. As the FIM and FIM+FAM are rated by team consensus, inter-rater reliability should be estimated by examining the agreement between ratings generated by different teams. As it was not possible to introduce this methodology into the routine clinical practice of each study unit, an alternative approach was used. The team consensus rating is compared with a rating made by the study co-ordinator at each unit. Study co-ordinators' ratings are based on information obtained from independent interviews with each member of the team before the team consensus rating was made.

Intra-rater and inter-rater reproducibility analyses are undertaken on subsamples of patients. Patients were allocated to either intra-rater or inter-rater reproducibility samples prior to entry into the study. All reproducibility analyses are reported as intraclass correlation coefficients, i.e., the percentage of total variance that results from true variance among patients (273, 279-282). Estimates of variance are obtained from repeated measures analysis of variance under random effects models (80). Lower limit confidence intervals are calculated according Fleiss' formula (80). Minimum recommended standards for reproducibility are .70 for group comparisons and .90 to .95 for individual comparisons (130, 271).

3a.1.2.3 Standard error of measurement (SEM)

Standard errors of measurement are reported because of the difficulty in interpreting reliability coefficients for individual patient scores (13). Measurement error introduces uncertainty about individual scores, the magnitude of which is indicated by the SEM (85).

Standard errors of measurement are estimates of the standard deviation of scores obtained with repeated administrations of an instrument to the same individual (217). Therefore, they are direct indicators of the probable extent of random error associated with scores and can be used to calculate confidence intervals around individual patient scores (85). Standard errors of measurement and 95% confidence intervals (95%CI) around true scores are calculated as follows:

$$SEM = SD \times \sqrt{(1 - reliability)}$$

$$95\%CI = true\ score +/- 1.96\ SEM$$

where: *SD* = standard deviation of admission scores;

reliability = a reliability coefficient (discussed further below).

It should be noted that confidence intervals are symmetrical around the true score for an individual but asymmetrical around their observed score. This is because scores tend to be biased, high scores tending to be biased upward and low scores downward. Estimates of unbiased (true) scores are the average scores that would be obtained if the instrument was administered on multiple occasions (217).

In this study 95% confidence intervals around individual scores are calculated using alphas and intra-rater reproducibility coefficients as the estimates for reliability. These reliability coefficients are used as each has a different interpretation. When confidence intervals are calculated from alpha coefficients they estimate cross-sectional measurement error at a single point in time and reflect instrument accuracy for individual patient assessment and clinical decision making (128). When confidence intervals are calculated from intra-rater reproducibility coefficients they estimate longitudinal measurement error and gauge the likelihood that an individual patient's change in score is attributable to random error rather than true change (266).

3a.1.3 Validity

An instrument is valid if it measures the construct it purports to measure (93). Whilst valid measurement of a physical parameter such as length is easy to verify, valid measurement of health constructs like disability is not an all-or-none property. Under these circumstances evidence must be gathered to determine the degree to which an instrument measures the construct it purports to measure.

There are three types of validity: content, criterion, and construct (93). Construct validity is the principal method used in this study. Content validity, the adequacy of item sampling, is not assessed as this is usually an aspect of instrument development. It is supported by appropriate methods of item generation and selection (271). Criterion-related validity, the degree to which a measure correlates with a gold standard (the criterion), is not assessed as gold standard measures of disability in patients with neurological disease do not exist (164).

Construct validity is the process used to establish the validity of a measurement instrument when no criterion or universe of content is accepted as entirely adequate to define the attribute being measured (283). Construct validity involves the generation of hypotheses concerning the construct the instrument is purported to measure and examination of the

extent to which empirical data support these hypotheses (283). Although there are several methods for determining construct validity, two categories have been distinguished by Bohrnstedt (14). He termed them internal and external construct validity. Internal construct validity involves statistical analyses of scale scores to determine if hypotheses concerning the theoretical structure of the instrument are supported. In contrast, external construct validity examines the relationships between scores on the instrument and other variables or measures to determine if hypotheses concerning the interpretation of scores are supported by empirical data. In this study evidence for both types of construct validity is examined. Other authors have categorised construct validity in a similar manner to Bohrnstedt and termed them psychometric and clinical test of validity (245), and logical and empirical analyses of validity (284).

3a.1.3.1 Internal construct validity

Two types of analyses are undertaken to examine the internal construct validity of FIM and FIM+FAM scales: internal consistency and intercorrelations between scales.

Evidence for the internal consistency of FIM and FIM+FAM scales also provides evidence for internal construct validity as it indicates the extent to which items are interrelated. In order to combine items to generate a score,

items should be homogeneous, that is, measure different aspects of the same attribute. Internal consistency is a necessary but not sufficient condition for homogeneity (unidimensionality) (285, 286). Construct validity is supported when minimum requirements for internal consistency, outlined above in the section on reliability, are satisfied.

Intercorrelations between scales of the FIM and FIM+FAM, as evaluated by Pearson's product-moment correlations, indicate the strength of relationships between the total, motor, and cognitive scales for each instrument. Construct validity is supported when the scales of an instrument are shown to be measuring related but different dimensions, and when the magnitude and pattern of the intercorrelations between scales conform to hypotheses based on theoretical considerations. Four specific predictions are tested concerning the intercorrelations between FIM and FIM+FAM scales. First, all intercorrelations between scales should be substantial (> .50) as all these scales measure different aspects of the same construct. Second, correlations between motor and cognitive scales of both instruments are expected to be in the .50 to .70 range, as these scales are purported to measure distinct subconstructs of disability. Third, because the total scales contain the motor and cognitive scales, correlations between the motor and cognitive scales of both instruments are predicted to be lower than correlations between each of these two scales and the total scales. Fourth, the correlation between the motor and total scales of the FIM is predicted to be higher than the correlation between the cognitive and

total scales. This is because of greater item overlap between the motor and total scales (13 items) than between the cognitive and total scales (5 items). For the FIM+FAM, the motor and cognitive scales are predicted to have similar correlations with the total scale as the degree of item overlap is similar (16 and 14 items respectively).

3b.1.3.2 External construct validity

Two types of analyses are undertaken to determine the external construct validity of FIM and FIM+FAM scales: correlations with other variables and known group discrimination.

Examining Pearson's product-moment correlations between FIM and FIM+FAM scales and other variables aims to provide evidence for convergent and discriminant validity. High correlations with measures of similar constructs indicates that an instrument measures the construct it is purported to measure and provides evidence of convergent validity. Low correlations with measures of dissimilar constructs indicates that an instrument does not measure a construct other than the one it is devised to measure and provides evidence of discriminant validity (283). Correlations are examined between FIM and FIM+FAM scales and other measures including: four measures of disability; measures of handicap, health status and psychological distress; seven measures of neuropsychological

functioning; and age. To the extent that observed associations agree with predicted relationships based on theoretical considerations, external construct validity is supported.

The expected direction, strength, and patterns of correlations between each FIM and FIM+FAM scale and other measures are presented in Table 3.1. In general, predictions are that greater disability will be associated with more handicap and poorer health status and psychological well-being. Similarly, greater cognitive disability will be associated with greater impairment as measured by neuropsychological tests.

As shown in Table 3.1, predictions concerning the strength and pattern of relationships between FIM and FIM+FAM scales and other measures are based on the conceptual similarity of the health constructs measured. They are defined approximately as strong ($r > .70$), moderate to substantial ($.30 < r < .70$), or weak ($r < .30$) (245). For example, the FIM total score is predicted to be highly correlated with other measures of disability, moderately correlated with measures of handicap and physical health status, and poorly correlated with measures of mental health status and psychological well-being. In addition, predictions are made concerning the order of magnitude of correlations between each validating instrument and the total, motor, and cognitive scales of the FIM and FIM+FAM (see right hand column of Table 3.1). For example, correlations with measures of

physical disability are predicted to be highest for the motor scales and lowest for the cognitive scales.

For each FIM and FIM+FAM scale, two types of evidence are examined to evaluate convergent validity. First, correlations for each FIM or FIM+FAM scale with measures of the same disability subconstruct should exceed correlations with all other measures. For example, the highest correlations for the FIM motor scale should be with other measures of physical disability. Second, the magnitude of correlations should conform to predictions as specified above. For example, correlations between the FIM motor scale and other measures of physical disability should exceed .70

Three types of evidence are examined to evaluate the discriminant validity of each FIM and FIM+FAM scale. First, each scale is expected to discriminate disability from other health constructs. That is, for all FIM and FIM+FAM scales correlations with measures of other related but different constructs such as handicap, health status, and psychological distress should be low to moderate. Furthermore, the pattern of these correlations should be consistent with predictions. For example, correlations with measures of handicap should exceed correlations with measures of psychological distress as disability is conceptually more similar to handicap. Second, FIM and FIM+FAM scores should be uncorrelated with age. Third, correlations among the FIM and FIM+FAM scales should exceed

correlations between these scales and measures of other health constructs (i.e. handicap, psychological distress and health status). This is because all FIM+FAM scales are measuring aspects of disability. Correlations between the cognitive scales of the FIM and FIM+FAM and measures of neuropsychological functioning are predicted to be substantial but not very high (>.80). This is because the cognitive scales of the FIM and FIM+FAM measure disability whereas measures of neuropsychological functioning measure individual cognitive impairments such as memory and reasoning.

Group differences (or known-groups) construct validity is supported when an instrument demonstrates the ability to detect differences in groups known or hypothesised to differ in the construct being measured (283). This is usually undertaken by the statistical comparison of mean scores for the groups of interest (287). In this study, the ability of FIM and FIM+FAM scores to discriminate between groups defined on the basis of staff-rated improvement and diagnostic category is tested.

The first hypothesis is that change scores for all FIM and FIM+FAM scales will be higher for groups defined as having undergone greater improvement in disability during neurorehabilitation. To test the hypothesis, FIM and FIM+FAM change scores are compared for patients whose level of improvement is rated by staff on a 4-point scale (1 = no change, 4 = marked improvement). External construct validity is supported when a stepwise

increase in the magnitude of FIM and FIM+FAM change scores parallels staff-rated improvement. The evidence for construct validity is stronger when these differences are statistically significant. Mean change scores for the different groups are compared using one-way analysis of variance with Duncan's multiple range tests for *post hoc* comparisons. Results are reported in terms of the magnitude and statistical significance of change scores as clinical and statistical significance are not necessarily equivalent (288, 289).

The second hypothesis is that stroke patients will demonstrate greater improvement after rehabilitation than MS patients as measured by all three FIM and FIM+FAM scales. This is because the overwhelming majority of MS patients studied have progressive disease, whereas stroke is an acute neurological event from which some recovery typically occurs. To test this hypothesis FIM and FIM+FAM change scores are compared for stroke and MS patients using independent sample *t* -tests.

3a.1.4 Responsiveness

Responsiveness is defined as the ability of an instrument to detect clinically significant change in the construct being measured (110). There is a debate as to whether responsiveness should be considered an aspect of validity (290, 291) or a separate psychometric property (112). Several methods

have been proposed for evaluating responsiveness, but there is no clear consensus as to which is the optimal method (292). Most methods examine scores at two points in time, usually before and after an intervention known to influence the construct being measured. Responsiveness is assessed on the basis of the magnitude of the standardised change score.

In this study FIM and FIM+FAM admission and discharge scores are compared. Responsiveness is determined by calculating an effect size statistic, of which there are several types (134, 291, 293). In this study, effect size is defined as the mean change score (admission minus discharge scores) divided by the standard deviation of the admission scores (132). Larger effect sizes indicate greater responsiveness. Values are interpreted using Cohen's criteria (small = .2, medium = .5, large = .8; 132, 293). In addition, the responsiveness of the FIM and FIM+FAM is compared with the Barthel Index, Modified Barthel Index, EDSS, and OPCS by comparing effect sizes.

3a.2 Comparison of the psychometric properties of the FIM and FIM+FAM in stroke and MS patients

The aim of these analyses is to compare the acceptability, reliability, validity, and responsiveness of FIM and FIM+FAM scales in patients with stroke and MS.

3a.2.1 Acceptability

Acceptability of FIM and FIM+FAM scales in stroke and MS patients is compared by examining score distributions. Criteria for acceptability are outlined above.

3a.2.2 Reliability

Reliability of FIM and FIM+FAM scales in stroke and MS patients is compared in terms of internal consistency, intra-rater and inter-rater reproducibility. Criteria for interpreting reliability coefficients are outlined above.

3a.2.3 Validity

Validity of FIM and FIM+FAM scales in stroke and MS patients is compared by examining the internal construct validity (internal consistency and intercorrelations between scales) and external construct validity (correlations with the Barthel Index, London Handicap Scale, SF-36, and age).

3a.2.4 Responsiveness

Responsiveness of FIM and FIM+FAM scales in stroke and MS patients is compared by examining effect sizes.

3a.3 Comparison of the psychometric properties of the FIM, FIM+FAM and Barthel Index

The aim of these analyses is to compare the acceptability, reliability, validity, and responsiveness of the FIM, FIM+FAM, and Barthel Index in order to determine the incremental validity (294) of each scale relative to the others. Although all three instruments are disability measures, they measure distinct aspects of disability. Consequently, measures of the same aspect of disability are compared: FIM and FIM+FAM total scales are compared as measures of global disability; FIM and FIM+FAM motor scales and the Barthel Index are compared as measures of motor disability; and FIM and FIM+FAM cognitive scales are compared as measures of cognitive disability.

3a.3.1 Acceptability

Acceptability is compared in terms of the percentage of missing data and descriptive statistics for scale scores for all three measures.

3a.3.2 Reliability

Reliability is compared in terms of internal consistency and reproducibility for all three measures. Reproducibility data are not available for the Barthel Index as this was not part of the original study protocol.

3a.3.3 Validity

Validity is compared in terms of internal and external construct validity for all three measures.

3a.3.3.1 Internal construct validity

Internal construct validity is compared by examining internal consistency. Internal consistency is compared on the basis of corrected item-total correlations, Cronbach's alphas, and homogeneity coefficients. Homogeneity coefficients are particularly useful for comparing the internal consistency of these three measures as alpha coefficients are influenced by their differing scale lengths.

3a.3.3.2 External construct validity

External construct validity is compared by examining concurrent validity, convergent and discriminant validity, and group differences validity. Concurrent validity is the extent to which the measure of a construct predicts (or is correlated with) another measure of the same construct evaluated at the same point in time (217). The extent to which scales measuring the same aspect of disability (e.g. FIM and FIM+FAM motor scales and the Barthel Index) predict each other is determined by the magnitude of the Pearson's product-moment correlation. Convergent and discriminant validity are compared by examining correlations between scales purporting to measure the same aspect of disability and other measures of disability, handicap, health status, psychological well-being, and neuropsychological functioning administered at the same time. Correlations with age are also compared. The extent of the similarity between the magnitude and pattern of these correlations for measures of the same aspect of disability indicate how similar they are with respect to their convergent and discriminant validity.

Group differences validity is compared by examining the relative measurement precision of comparable scales for all three measures. Measurement precision is the extent to which an instrument can detect small differences in the construct being measured (36). Using the group differences method of examining validity, the measurement precision of an

instrument is defined as the degree to which it separates the groups relative to within-groups variance (100). The *F* -statistic, derived from a one-way analysis of variance, takes both of these attributes into account as it defines the ratio of between-groups (systematic) variance to within-group (error) variance. Therefore, the higher the *F* -statistic the greater the measurement precision (245). By comparing different instruments in the same sample, relative measurement precision can be estimated by the ratio of pairwise *F* -statistics (*F* for one measure divided by *F* for another). Consequently, relative measurement precision estimates, in proportional terms, how much more (or less) precise one measure is compared to another in detecting group differences (100).

The importance of measurement precision lies in the trade-off between sample size and statistical power. This fact has implications for clinical trials as better measurement precision in detecting group differences means higher power for a fixed sample size or fewer patients to achieve a fixed level of statistical power (133). That is, for a given sample size the greater the measurement precision of an instrument the more likely it is to demonstrate statistically significant results. Alternatively, using instruments with better measurement precision in detecting group differences means that smaller samples can be used to detect statistically significant results (133).

The relative measurement precision of the FIM, FIM+FAM, and Barthel Index is evaluated by comparing their ability to discriminate between patients on the basis of different levels of staff-rated improvement in disability. Improvement in disability from admission to discharge for each patient is rated by staff as minimal, moderate, or marked. For each of these three groups (minimal, moderate or marked improvement) mean change scores are calculated for the FIM, FIM+FAM, and the Barthel Index. Mean change scores for each scale are compared using one-way analysis of variance to generate an *F* -statistic. From each group of scales being compared (total, motor, or cognitive scales), one measure is chosen arbitrarily as the standard and the others are compared using pairwise *F*-statistics: i.e. *F* -statistic for the scale of interest divided by *F* -statistic for the arbitrary standard. The relevant FIM scale in each group of scales being compared is used as the arbitrary standard and, therefore, will be assigned a relative measurement precision of 1. For other scales, values > 1 (or < 1) indicate in percentage terms greater (or lesser) measurement precision than the comparable FIM scale.

3a.3.4 Responsiveness

Responsiveness is compared on the basis of effect sizes. For measures of the same aspect of disability, the instrument with the largest effect size is considered the most responsive.

3b Results

Results are presented in five sections followed by a summary. The first section describes the patient samples. The next two sections present results of analyses of the psychometric properties of the FIM and FIM+FAM, respectively. The fourth section presents results of analyses comparing the psychometric properties of the FIM and FIM+FAM in stroke and MS patients. The final section presents results of analyses comparing the psychometric properties of comparable scales of the FIM, FIM+FAM, and Barthel Index.

3b.1 Patient samples

3b.1.1 Descriptive characteristics of patient samples

A total of 214 neurologically disabled inpatients from three sites were invited to participate in the study. Two patients declined to participate: both were from the RRU and no specific reasons were given. Three of the 212 patients who agreed to participate in the study were discharged within one week of admission and were, therefore, withdrawn for failing to meet inclusion criteria. One patient self-discharged, and two patients were transferred due to clinical deterioration; all three patients were from NRU. The final study sample consisted of the remaining 209 patients.

Table 3.2 presents characteristics of patients in the total sample and at the three clinical sites compared to the 1994 local population treated between 1st April 1994 and 31st March 1995. Men and women are evenly represented and patients across a wide range of ages are included. Stroke, MS, and head injury patients are the three largest diagnostic groups, constituting 80.4% of the sample. The remaining 19.6% of the total sample represents a mix of neurological disorders.

The mean length of stay for patients in the total sample was approximately nine weeks (range 11 days to 14 months). Length of stay calculations only include patients who completed their planned in-patient rehabilitation programme (n = 194; 92.8% of sample). Of the 15 patients who did not complete their planned rehabilitation, seven were still in-patients at the end of data collection, six were transferred for acute hospitalisation, one self-discharged, and one was transferred to another rehabilitation unit for ongoing rehabilitation.

The three clinical sites differ in terms of numbers of patients enrolled, proportion of men, casemix, and length of stay. The NRU has the largest sample, a large proportion of MS patients, and the shortest length of stay. These findings reflect the longer duration of recruitment at the NRU and its expertise in short stay rehabilitation of MS patients. The RNRU has the

highest proportion of head injury patients and of men and the longest length of stay. The RRU has the highest proportion of stroke patients and of women and a medium length of stay.

Samples from the three clinical sites are largely representative of their respective 1994 populations. The NRU sample has more MS patients and fewer patients with other diagnoses than the local population. The RNRU sample is similar to the 1994 local population as almost all patients were enrolled into the study. The RRU has fewer males and head injury patients, more stroke patients, and fewer patients in the other three diagnostic categories compared with the 1994 local population. The small size of the RRU sample and the fact that most patients with head injuries are male are likely to have contributed to these differences.

3b.1.2 Characteristics of patients with stroke and MS

Table 3.3 presents characteristics of patients with stroke and MS which form the two largest diagnostic groups in the study sample. Differences between the two groups are compared qualitatively rather than statistically as they are expected to differ considerably. Consistent with clinical expectation, the two diagnostic groups differ in gender, age, and length of stay, and are not distributed equally among the three clinical sites. Multiple sclerosis predominantly affects young women, whereas stroke tends to affect men

and women more equally, especially in the younger age group. Length of stay for MS patients in this study is short as most of these patients have progressive disabilities and rehabilitation is highly focused towards specific aims. In addition, all MS patients are at the NRU where short stay rehabilitation is a feature. In contrast, stroke patients in this study have longer stays reflecting the impact of acute disability and the potential for gradual recovery.

In the stroke subgroup, 72% have unilateral hemispheric lesions, 20% have subarachnoid haemorrhage, 6 % have brainstem stroke, and 2% have bilateral hemisphere stroke. Of the hemispheric strokes, 73% are infarcts and 59% involve the dominant hemisphere. In the MS subgroup, the disease type is secondary progressive in 81%, primary progressive in 11%, and relapsing remitting disease in 8%. These figures are similar to population statistics for stroke (295) but not for MS reflecting the greater disability that is associated with the progressive form of MS.

3b.1.3 Characteristics of patients in the intra- and inter-rater reproducibility subsamples

Table 3.4 presents a comparison of FIM and FIM+FAM admission scores for patients in the intra- and inter-rater reproducibility subsamples with patients in the total sample. There are no statistically significant differences between

either subsample and the total sample. These results indicate that the samples used to assess intra and inter-rater reproducibility are representative of the total sample.

3b.2 Psychometric evaluation of the FIM

3b.2.1 Acceptability

The fact that there are no missing FIM data indicate that the instrument has been successfully incorporated into clinical practice at the three study sites. Table 3.5 presents descriptive statistics and floor and ceiling effects for FIM items at admission. Responses are well distributed across response categories. Only one response category for one item (feeding - response 3) is not endorsed by any patient. Maximum endorsement frequencies range from 19.1% for problem solving to 56.0% for stairs (mean 36.0%). Item floor effects range from 7.2% for comprehension to 56.0% for stairs (mean 19.4%). Item ceiling effects range from 1.9% for shower/tub transfer to 47.4% for comprehension (mean 26.2%). Maximum endorsement frequencies, floor effects, and ceiling effects do not exceed the maximum criterion of 80% used in this study. These results indicate a relatively even distribution of disabilities, as defined by FIM items, in the study sample.

Table 3.6 presents descriptive statistics, floor and ceiling effects, and skewness statistics for FIM scale scores at admission. Scores span almost the entire range with mean scores near to and consistently higher than the scale mid-point. Floor and ceiling effects are generally small and none exceed the maximum recommended criterion of 15%. All skewness statistics are within the recommended range of -1 to +1. These results indicate that FIM scales adequately represent the range of severity of disabilities in patients in this sample, and demonstrate good variability in the constructs they measure.

3b.2.2 Reliability

3b.2.2.1 Internal consistency

Table 3.7 presents reliability estimates for FIM scales. High internal consistency is demonstrated for all three FIM scales. Item-total correlations exceed the required minimum standard of .30, indicating that all items are substantially linearly related to scale scores. Alpha coefficients exceed .90 for all scales, indicating that each scale satisfies minimum internal consistency criteria for both group and individual comparisons. When items are deleted, alpha coefficients do not increase substantially (results not shown), indicating that no individual items compromise the internal consistency of FIM scales. Homogeneity coefficients exceed .30 and, therefore, satisfy minimum requirements.

3b.2.2.2 Reproducibility

High reproducibility is also demonstrated for all three FIM scales. Intraclass correlation coefficients for both intra- and inter-rater reproducibility exceed .90, indicating a high degree of stability of FIM scores over time and agreement between raters. Minimum standards for individual and group comparisons are satisfied.

Despite the high internal consistency and reproducibility reported above, the estimated 95% confidence intervals for individual scores are large. For example, the cross-sectional 95% confidence interval for an individual patient's FIM total score is +/- 11.5 points. This confidence interval of 23 points on the FIM comprises 21.3% of the possible score range and indicates that individual scores must be interpreted with caution. Similarly, the longitudinal confidence intervals indicate that FIM total scores for an individual patient will have to change by at least 8.1 points for the change to be considered statistically significant. In other words, there is a high probability that FIM total change scores of 8 points or less are due to random error.

3b.2.3 Validity

3b.2.3.1 Internal construct validity

Evidence for the internal consistency of FIM scales (see Table 3.7) also supports their construct validity as high internal consistency supports scale homogeneity. As predicted, item-total correlations and homogeneity coefficients are higher for motor and cognitive scales than for the total scale. These results provide evidence to support two distinct subconstructs, motor and cognitive disability, within the overall construct of disability.

Table 3.8 presents intercorrelations between the three FIM scales. Intercorrelations are moderate to high, indicating that the three scales are measuring related constructs and thus providing evidence for convergent validity. As predicted, the correlation between the motor and cognitive scales is in the range .50 to .70 and is lower than the correlations between each of these two scales and the total scale. These results indicate that the motor and cognitive scales measure related but separate constructs and provide evidence for discriminant validity. As predicted the correlation between the motor and total scale exceeds the correlation between the cognitive and total scale. The correlation between the motor and total scales is very high. This raises a concern about whether the motor and total scales indeed measure distinct constructs. However, the correlations between the motor and total scales with the cognitive scale are quite different, indicating

that the motor and total scales are measuring related but different constructs.

3b.2.3.2 External construct validity

Table 3.9 presents correlations between FIM scales and measures of disability, handicap, health status, psychological distress, and age. Table 3.11 presents correlations between FIM scales and neuropsychological measures.

The FIM total scale is highly correlated with the four disability measures and moderately correlated with neuropsychological measures and measures of handicap and physical health status. Correlations with mental health status and psychological distress are low. The direction, magnitude, and pattern of these correlations are consistent with predictions and provide evidence for the convergent and discriminant validity of the FIM total scale as a measure of global disability.

The FIM motor scale is also highly correlated with all four measures of disability, and is moderately correlated with neuropsychological measures and measures of handicap and physical health status. Correlations with mental health status and psychological distress are low. Correlations

between the FIM motor scale and neuropsychological measures are lower than correlations between the FIM total scale and neuropsychological measures. Correlations with measures of disability, handicap, and physical health status are marginally but consistently higher for the FIM motor than total scale. These findings provide evidence for the convergent and discriminant validity of the FIM motor scale as a measure of motor disability.

Correlations between the FIM cognitive scale and other disability measures are generally moderate and lower than those found for the FIM total and motor scales. Correlations between the FIM cognitive scale and neuropsychological measures are moderate or high, and generally higher than those for either the total or motor scale. Correlations with handicap, physical and mental health status and psychological distress are low. Unlike the FIM total and motor scales, the cognitive scale has low correlations with handicap and physical health status. These findings provide evidence for the convergent and discriminant validity of the FIM cognitive scale as a measure of cognitive disability.

Finally, all three FIM scales are uncorrelated with age. This finding provides further evidence of discriminant validity for the FIM.

Tables 3.11 and 3.12 provide evidence for group differences validity of FIM scales. Table 3.11 presents mean FIM change scores associated with different levels of staff-rated improvement in disability. Staff ratings of change in disability are available for 181 patients: data for 13 patients are missing and two persons whose disability worsened were excluded. Results are presented for the remaining 179 patients. Negative change scores indicate improvement in disability at discharge. As hypothesised, a stepwise pattern of improvement in disability as measured by all three FIM scales is associated with statistically significant improvement in disability as assessed by staff ratings.

Post hoc comparisons reveal that for all FIM scales, change scores for patients rated as markedly improved are significantly higher ($p < .05$) than for patients rated as moderately improved. In addition, change scores for patients rated as moderately improved are significantly higher ($p < .05$) than for patients rated as minimally improved. However, there are no significant differences ($p > .05$) in FIM change scores between patients rated as minimally improved and those rated as showing no improvement. These results provide evidence that FIM scores are able to detect differences between groups distinguished by staff-rated improvements in disability.

Table 3.12 presents FIM change scores for stroke and MS patients. Change scores for FIM total and motor scales are significantly higher for stroke than

for MS patients, but do not differ for cognitive scale scores, indicating greater improvement in stroke patients. These results provide evidence that FIM scores are able to detect differences between groups distinguished on the basis of diagnosis.

3b.2.4 Responsiveness

Table 3.13 presents FIM admission, discharge, and change scores with responsiveness reported as effect sizes. All change scores are statistically significant ($p < 0.001$). Effect sizes indicate that the responsiveness of the FIM motor and total scales is medium, whilst the responsiveness of the cognitive scale is small.

Table 3.14 shows the relative responsiveness of the FIM compared with other disability measures. Effect size calculations show that the FIM motor and total scales, FIM+FAM motor and total scales, Modified Barthel Index, and Barthel Index have similar medium responsiveness. The responsiveness of the OPCS is small to medium. The responsiveness of the cognitive scales of the FIM and FIM+FAM is similar and small. The responsiveness of the EDSS is very small.

3b.3 Psychometric evaluation of the FIM+FAM

3b.3.1 Acceptability

There are no missing FIM+FAM data indicating that the instrument has been successfully incorporated into clinical practice at all three study sites. Table 3.15 presents descriptive statistics and floor and ceiling effects for FIM+FAM items at admission. Responses are well distributed across response categories. Only one response category for one item (feeding - response 3) is not endorsed by any patient. Maximum endorsement frequencies range from 18.7% for adjustment to limitations to 70.8% for swallowing (mean 37.2%). Item floor effects range from 3.3% for swallowing to 56.0% for stairs (mean 17.7%). Item ceiling effects range from 1.9% for shower/tub transfer to 70.8% for swallowing (mean 28.4%). Maximum endorsement frequencies and floor and ceiling effects do not exceed the maximum criterion of 80% used in this study. These results indicate a relatively distribution of disabilities, as defined by the FIM+FAM items, in the study sample.

Table 3.16 presents descriptive statistics, floor and ceiling effects, and skewness statistics for FIM+FAM scales at admission. Scores span almost the entire possible ranges, with mean scores near to and but consistently higher than the scale mid-point. Floor and ceiling effects are small, ranging from 0% to 1.9%. Skewness statistics are within the recommended range of -1 to +1. These results indicate that FIM+FAM scales adequately represent

the range of severity of disabilities in the study sample and demonstrate good variability in the constructs they measure.

3b.3.2 Reliability

Table 3.17 presents reliability estimates for FIM+FAM scales. High internal consistency and reproducibility are demonstrated for all three scales. Item-total correlations and homogeneity coefficients exceed minimum requirements. Alpha coefficients and intraclass correlation coefficients for intra- and inter-rater reproducibility are very high (all approach 1.00), exceeding minimum requirements for individual and group comparison studies. Confidence intervals for individual scores are large. For example, the cross-sectional 95% confidence band for FIM+FAM total scores is 31 points, 17.2% of the score range, indicating that individual patient scores should be interpreted cautiously. Similarly, longitudinal 95% confidence intervals indicate that a FIM+FAM total score for an individual should change by more than 12.6 points to be considered statistically significant.

3b.3.3 Validity

3b.3.3.1 Internal construct validity

Evidence for the internal consistency of FIM+FAM scales (see Table 3.17) supports their internal construct validity as high internal consistency supports scale homogeneity. Item-total correlations and homogeneity coefficients are higher for the motor and cognitive scales than for the total scale. These findings provide evidence to support the existence of motor and cognitive disability subconstructs within an overall construct of disability.

Table 3.18 presents intercorrelations between FIM+FAM scales. All correlations are high, indicating that FIM+FAM scales are measuring related constructs. These findings provide evidence for convergent validity. The correlation between motor and cognitive scales is lower than correlations between each of these scales and the total scale, indicating that these scales are measuring related but separate constructs and providing evidence of discriminant validity. Very high correlations between motor and total scales (.93) and between cognitive and total scales (.91) suggests that these three scales are measuring the same construct. However, different correlations between motor and total with cognitive scales (.69 and .91), and cognitive and total with motor scales (.69 and .93) indicate that the scales are measuring related but different constructs.

3b.3.3.2 External construct validity

Table 3.19 presents correlations between FIM+FAM scales and measures of disability, handicap, health status, psychological distress, and age. Table 3.20 presents correlations between FIM+FAM scales and neuropsychological measures.

The FIM+FAM total scale is highly correlated with all four disability measures, and moderately correlated with neuropsychological measures and measures of handicap and physical health status. Correlations with mental health status and psychological distress are low. The direction, magnitude and pattern of these correlations are consistent with predictions and provide evidence for the convergent and discriminant validity of the FIM+FAM total scale as a measure of global disability.

The FIM+FAM motor scale is also highly correlated with all four measures of disability, and is moderately correlated with neuropsychological measures and measures of handicap and physical health status. Correlations with mental health status and psychological distress are low. Correlations between the FIM+FAM motor scale and neuropsychological measures are lower than correlations between the FIM+FAM total scale and neuropsychological measures. Correlations with measures of disability, handicap, and physical health status are higher for the FIM+FAM motor

than total scale. These findings provide evidence for the convergent and discriminant validity of the FIM motor scale as a measure of motor disability.

Correlations between the FIM+FAM cognitive scale and the four disability measures are generally moderate and lower than those found for the FIM+FAM total and motor scales. Correlations between the FIM+FAM cognitive scale and neuropsychological measures are moderate or high, and generally higher than those for either the total or motor scale. Correlations with handicap, physical and mental health status, and psychological distress are low. Unlike the FIM+FAM total and motor scales, the FIM+FAM cognitive scale shows a low correlation with handicap and physical health status. These findings provide evidence for the convergent and discriminant validity of the FIM+FAM cognitive scale as a measure of cognitive disability.

Finally, all three FIM+FAM scales are uncorrelated with age, providing further evidence of discriminant validity.

Tables 3.21 and 3.22 provide evidence for group differences validity of FIM+FAM scales. As for the FIM, staff ratings of change in disability are presented for 179 patients. Table 3.21 presents mean FIM+FAM change scores associated with different levels of staff-rated improvement in

disability. As hypothesised, a stepwise pattern of improvement in disability as measured by all three FIM+FAM scales is significantly associated with a similar pattern of improvement in disability as assessed by staff ratings.

Post hoc comparisons of mean scores reveal that for all FIM+FAM scales, change scores for patients rated as markedly improved are significantly higher ($p < .05$) than for patients rated as moderately improved. Also, change scores for patients rated as moderately improved are significantly higher than for patients rated as minimally improved. There are no significant differences ($p > .05$) in FIM+FAM change scores between patients rated as minimally improved and those rated as showing no improvement.

Table 3.22 presents FIM+FAM change scores for stroke and MS patients. Change scores for FIM+FAM total and motor scales are significantly higher for stroke than MS patients, but do not differ for cognitive scale scores, indicating greater improvement in stroke patients. These findings provide evidence that the FIM+FAM is able to detect differences between groups distinguished on the basis of diagnosis.

3b.3.4 Responsiveness

Table 3.23 presents FIM+FAM admission, discharge, and change scores with responsiveness reported as effect sizes. Change scores for all three scales indicate statistically significant improvements in disability from admission to discharge ($p < 0.001$). Effect sizes indicate medium responsiveness for the FIM+FAM motor and total scales and small responsiveness for the cognitive scale.

Table 3.14 shows the relative responsiveness of FIM+FAM scales compared to seven other measures. Effect size calculations show that the FIM+FAM motor and total scales, FIM motor and total scales, Modified Barthel Index, and Barthel Index have similar medium responsiveness. The responsiveness of the OPCS is small to medium. The responsiveness of the cognitive scales of the FIM+FAM and FIM is similar and small. The responsiveness of the EDSS is very small.

3b.4 Comparison of the psychometric properties of the FIM and FIM+FAM in stroke and MS patients

3b.4.1 FIM

Tables 3.24 and 3.25 present results comparing the psychometric properties of each of the three FIM scales in stroke and MS patients. Table 3.24 presents results for acceptability, reliability, and internal construct validity (internal consistency and intercorrelations between FIM scales). Table 3.25 presents results for external construct validity and responsiveness.

Table 3.24 shows that the acceptability, reliability, and internal construct validity of FIM scales are similar in stroke and MS patients. Some small differences in acceptability and internal consistency are demonstrated for the FIM cognitive scale. Unlike scores for stroke patients, scores for MS patients do not span the lower (more disabled) end of the scale and show a ceiling effect at the recommended upper limit of 15%. These findings indicate that the FIM cognitive scale is marginally more acceptable in stroke than MS patients.

Intercorrelations between scales are almost identical for stroke and MS patients, indicating that the strength of relationships between the three FIM scales is similar in the two patient groups. This finding, along with the

demonstration of similar internal consistency in Table 3.24, indicate similar internal construct validity for all FIM scales in stroke and MS patients.

Table 3.25 shows that correlations between FIM scales and the Barthel Index, SF-36 PCS and MCS, and age are comparable for stroke and MS patients, indicating similar convergent and discriminant construct validity. However, correlations between the FIM scales and the London Handicap Scale are higher for MS than for stroke patients. These results indicate that the FIM is less able to discriminate between disability and handicap in MS than in stroke patients.

Table 3.25 also demonstrates that all three FIM scales have larger effect sizes for stroke than for MS patients indicating that all FIM scales are more responsive in stroke than in MS patients. In both disease groups the cognitive scales are less responsive than the motor and total scales.

3b.4.2 FIM+FAM

Tables 3.26 and 3.27 present results comparing the psychometric properties of the three FIM+FAM scales in stroke and MS patients. Table 3.26 presents results for acceptability, reliability, and internal construct validity.

Table 3.27 presents results for external construct validity and responsiveness.

Table 3.26 shows similar acceptability, reliability, and internal construct validity for all three scales in both patient groups. One small difference in acceptability between the two diseases is that FIM+FAM cognitive scores for MS patients do not span the lower (more disabled) end of the scale range. However, this is not associated with a notable ceiling effect. Intercorrelations between scales are almost identical for the two patient groups. These findings, together with the demonstration of similar internal consistency in Table 3.26, indicate similar internal construct validity of FIM+FAM scales in stroke and MS patients.

Table 3.27 shows that for all three FIM+FAM scales, correlations between FIM+FAM scales and the Barthel Index, SF-36 PCS and MCS, and age are comparable in stroke and MS patients, indicating similar convergent and discriminant validity. However, correlations between FIM+FAM scales and the London Handicap Scale are higher for MS than stroke patients indicating that the FIM+FAM is less able to discriminate between disability and handicap in MS than in stroke patients. Table 3.27 also shows that all three FIM+FAM scales have larger effect sizes for stroke than for MS patients, indicating greater responsiveness in stroke than in MS patients.

3b.5 Comparison of the psychometric properties of the FIM, FIM+FAM, and Barthel Index

The Barthel Index was administered at two study sites, the NRU and RRU. Therefore, FIM, FIM+FAM, and Barthel Index scores are available for 149 patients. Tables 3.28 to 3.34 inclusive present results comparing the psychometric properties of the three instruments. Each table is arranged such that scales measuring the same aspects of disability are group together: FIM and FIM+FAM total scales as measures of global disability; FIM and FIM+FAM motor scales and the Barthel Index as measures of motor disability; FIM and FIM+FAM cognitive scales as measures of cognitive disability.

3b.5.1 Acceptability

There are no missing data for any of the three measures. Table 3.28 presents score ranges, means and standard deviations, floor and ceiling effects, and skewness statistics for the FIM, FIM+FAM, and Barthel Index. Descriptive statistics are similar for both global disability measures and all criteria of acceptability are satisfied. These findings indicate no advantage in acceptability for either FIM or FIM+FAM total scales. Similar findings of comparable acceptability are demonstrated for the three measures of motor disability. Ceiling effects and skewness for the FIM cognitive scale exceed recommended criteria only slightly. These results suggest small advantages

in terms of acceptability for the FIM+FAM cognitive scale over the FIM cognitive scale as a measure of cognitive disability.

3b.5.2 Reliability

Table 3.29 compares reliability estimates for the FIM, FIM+FAM, and Barthel Index. For all three aspects of disability measurement , the three instruments have nearly identical internal consistency and reproducibility. These results indicate that none of the three instruments has better reliability despite differences in the number of items in each scale. All scales satisfy minimum recommended reliability criteria for individual and group comparisons.

3b.3 Validity

3b.5.3.1 Internal construct validity

Internal consistency estimates for measures of global, motor, and cognitive disability are presented in Table 3.29. They are discussed above and provide evidence of similar internal construct validity for the different instruments.

3b.5.3.2 External construct validity

Tables 3.30 to 3.33 inclusive compare the external construct validity of scales of the FIM, FIM+FAM, and Barthel scales measuring global, motor, and cognitive disability. Table 3.30 presents intercorrelations between FIM and FIM+FAM scales and the Barthel Index. This table demonstrates two findings. First, correlations between measures of the same aspect of disability (in bold type) are very high. For example, intercorrelations between the FIM motor scale, FIM+FAM motor scale, and the Barthel Index which all measure motor disability range from .97 to .996. These results provide strong evidence for the concurrent validity of the two scales of global disability, for the three scales of motor disability, and for the two scales of cognitive disability. Second, scales measuring the same aspect of disability have very similar correlations with scales measuring other aspects of disability. For example, the FIM and FIM+FAM total scales have similar correlations with measures of motor disability and with measures of cognitive disability. These results provide evidence for similar convergent and discriminant validity of scales measuring the same aspects of disability.

Tables 3.31 and 3.32 provide further evidence that scales of the FIM, FIM+FAM and Barthel Index which measure the same aspects of disability have similar convergent and discriminant validity. Table 3.31 demonstrates near identical correlations between measures of global, motor, and cognitive disability and other measures of disability, handicap, health

status, psychological distress, and age. Table 3.32 demonstrates near identical correlations between measures of global, motor, and cognitive disability and measures of neuropsychological functioning.

Table 3.33 presents relative precision estimates for the FIM, FIM+FAM, and Barthel Index. All scales demonstrate a stepwise increase in change scores associated with increasing staff-rated improvement in disability. Whilst these results are statistically significant, there are notable differences in relative measurement precision. The FIM+FAM total scale is 81% as precise as the FIM total score. These results indicate that the FIM total scale is superior to the FIM+FAM total scale in discriminating between clinical groups differing in staff-rated improvement in disability. The FIM and FIM+FAM motor scales show almost identical measurement precision. However, the Barthel Index is only 61% as precise as the FIM motor scale, indicating that the FIM and FIM+FAM motor scales are superior for detecting group differences in motor disability. For the measurement of cognitive disability, the FIM and FIM+FAM cognitive scales have almost identical measurement precision in detecting group differences.

3b.5.4 Responsiveness

Table 3.34 presents admission, discharge, and change scores and effect sizes for scales of the FIM, FIM+FAM and Barthel Index measuring global,

motor, and cognitive disability. Data are available for 136 patients. Seven patients were still inpatients when data collection stopped, five patients acutely deteriorated during the rehabilitation period and were transferred elsewhere, and one patient was transferred to another unit for ongoing rehabilitation. Effect sizes indicate comparable responsiveness for scales measuring the same aspects of disability. These results suggest no advantage in responsiveness among the three measures.

3b.6 Summary of results

The FIM and FIM+FAM are acceptable, reliable, valid, and responsive measures of disability in the sample studied. The FIM and FIM+FAM are rigorous disability measures in stroke and MS patients. There are three psychometric differences between the two patient groups. The FIM cognitive scale may be less acceptable in MS than in stroke patients. All scales are more responsive in stroke than in MS suggesting that responsiveness may be disease-dependent. Correlations with handicap are higher for MS than for stroke.

The FIM, FIM+FAM, and Barthel Index have very similar measurement properties in patients undergoing inpatient neurorehabilitation. Comparable scales of the FIM and FIM+FAM have virtually identical psychometric properties. The Barthel Index has virtually identical psychometric properties

to the motor scales of the FIM and FIM+FAM, except for measurement precision. The total and motor scales of the FIM and FIM+FAM and the Barthel Index have very similar psychometric properties.

Chapter 4

Evaluating Conceptual Models Using Item Analysis: Method and Results

Chapter 3 provides evidence that the FIM and FIM+FAM are reliable, valid, and responsive measures of disability. However, when the FIM and FIM+FAM are compared with each other and with the Barthel Index, comparable scales of the three instruments are shown to have almost identical measurement properties. These findings indicate that the purported conceptual advantages of the FIM over the Barthel Index and the FIM+FAM over the FIM are not supported by empirical data. In addition, the finding of identical measurement properties of scales which differ in length suggests item redundancy in the FIM and FIM+FAM. The fact that the development of the FIM and FIM+FAM did not include an item reduction stage or a test of their underlying conceptual models raises two questions: to what extent are the conceptual models of the FIM and FIM+FAM supported by empirical data and, if there is item redundancy, can a short-form measure be developed?

This chapter addresses these two questions. A detailed item analysis is performed to determine the extent of empirical support for the conceptual

models of the FIM and FIM+FAM. Item reduction is then performed to determine whether a reliable and valid short-from measure can be developed.

4a Item analysis of the FIM and FIM+FAM

4a.1 Conceptual and measurement models of the FIM and FIM+FAM

The conceptual model of disability on which the FIM is based hypothesises that disability consists of two distinct domains, motor and cognitive disability, and that each domain consists of two or more subconstructs. For the motor domain, the subconstructs are self-care, sphincter care, transfer, and mobility. For the cognitive domain, the subconstructs are communication and social cognition.

The measurement model of an instrument refers to the scale and subscale structure and the procedures followed to create scale and subscale scores (130). Table 4.1 shows the measurement model for the FIM. The 18 items comprise two scales, a motor scale containing 13 items and a cognitive scale containing 5 items. Each of the two FIM scales consists of two or

more subscales. The motor scale has four subscales: self-care (6 items), sphincter care (2 items), transfer (3 items), and locomotion (2 items). The cognitive scale has two subscales: communication (2 items) and social cognition (3 items).

The conceptual model of disability on which the FIM+FAM is based is an extension of the conceptual model of the FIM. There are still motor and cognitive domains, but each domain consists of three or more subconstructs. For the motor domain, the subconstructs are self-care, sphincter care, transfer, and mobility. For the cognitive domain, the subconstructs are communication psychosocial adjustment, and cognitive function.

Table 4.2 shows the measurement model for the FIM+FAM. The 30 items comprise two scales, a motor scale containing 16 items and a cognitive scale containing 14 items. Each of the two FIM+FAM scales consists of three or more subscales. The motor scale has four subscales: self-care (7 items), sphincter care (2 items), transfer (4 items), and locomotion (3 items). The cognitive scale has three subscales: communication (5 items), psychosocial adjustment (4 items), and cognitive function (5 items).

Each FIM and FIM+FAM scale and subscale contains multiple items which are summed in accordance with Likert's method of summated ratings, without weighting, to generate total scores (296). The grouping of items into subscales and scales and the calculation of summated scores is based on four scaling assumptions. First, items within each scale or subscale measure the same construct. Second, items can be summed, without weighting, to generate scale and subscale scores. Third, scales and subscales measure distinct constructs. Fourth, the subscales and scales defined by the developers are the most appropriate method of grouping the items.

The following section evaluates each of these four scaling assumptions for the FIM and FIM+FAM. The method and results for each assumption are presented together.

4a.2 Do the items of each scale and subscale measure the same construct ?

4a.2.1 Method

Items measuring different aspects of the same underlying construct are inter-related (internally consistent). The extent to which items are internally consistent is determined by examining the results of three analyses:

correlations between all possible pairs of items (item intercorrelations); correlations between each item and the total scale or subscale score corrected for item overlap (corrected item-total correlation); and alpha coefficients for each scale and subscale.

Item intercorrelations indicate the extent to which items are related. It is recommended that the mean item-intercorrelation for a scale or subscale, also called the homogeneity coefficient (265), should exceed .30 (261, 275, 276). However, items that are highly correlated indicate that one item may be redundant. Although there is no widely accepted criterion for item redundancy, Juniper *et al.* (297) recommend that item-intercorrelations exceeding .70 indicate that an item can be removed.

Item-total correlations indicate the strength of relationship between individual items and the construct being measured. It is recommended that item-total correlations corrected for overlap should exceed .30 (217).

Alpha coefficients indicate the extent to which items in a scale are interrelated by comparing the variance of the total score to the sum of the variances of the individual items. As correlations between items increase, the variance for the total score increases and alpha coefficients approximate unity (269). It is recommended that alpha coefficients should exceed .70

(217). However, alpha coefficients exceeding .90 to .95 (286, 298), especially when there are less than 10 items (98, 299), suggest item redundancy.

4a.2.2 Results

Tables 4.3 and 4.4 present internal consistency estimates for scales and subscales of the FIM and FIM+FAM. Recommended criteria are satisfied for item-intercorrelations, item-total correlations, and alpha coefficients. These findings indicate that scales and subscales of both instruments are internally consistent and, therefore, their items measure the same construct.

However, all scales and most subscales of the FIM and FIM+FAM have item-intercorrelations greater than .70 and alpha coefficients which exceed .90. These results suggest item redundancy in the scales and subscales of both instruments.

4a.3 Can unweighted item scores be summed to generate scale and subscale scores ?

4a.3.1 Method

Likert proposed that items can be summed without weighting to generate scores when they: are substantially linearly related to the total score computed from all other items in that group; measure at similar points on the scale; contribute equally to the total score variance; and contribute equal proportions of information to the total score. These four criteria are satisfied when items are internally consistent and have similar mean scores, variances, and item-total correlations (296, 300, 301). However, when item-total correlations exceed .30, the criteria of equivalent item means, variances, and item-total correlations can be considered satisfied, even if they vary (302).

4a.3.2 Results

The first column of results in Tables 4.3 and 4.4 shows item-total correlations for scales and subscales of the FIM and FIM+FAM. All values are high (minimum values: FIM = .60; FIM+FAM = .55), indicating that Likert's criteria of internal consistency and equivalence of item means, variances, and item-total correlations can be considered satisfied. These

findings indicate that it is legitimate to sum unweighted item scores to generate scale and subscale scores.

4a.4 Do scales and subscales measure distinct constructs ?

4a.4.1 Method

To evaluate the extent to which FIM and FIM+FAM scales and subscales measure distinct dimensions of disability, product-moment correlations among scales and subscales are compared with reliability coefficients. The extent to which intercorrelations between scales are lower than their reliability estimates indicates the degree of unique reliable variance measured by each scale relative to the other (302). This is because the internal consistency reliability coefficient of a scale, its alpha coefficient, can be thought of as the correlation between a scale and itself (88). Therefore, when intercorrelations between scales are similar to their reliability, there is no evidence of unique reliable variance.

Three groups of correlations address the question of the distinctiveness of constructs measured by scales and subscales. Intercorrelations among scales indicate the extent to which scales are measuring distinct disability constructs; correlations between scales and subscales indicate the extent to which subscales are measuring constructs distinct from scales; and

intercorrelations among subscales indicate the extent to which independent item groups are measuring distinct aspects of disability. When there is item overlap, for example between the motor and total scales, correlations are predicted to be substantial. Although Nunnally and Bernstein (271) suggest that correlations among subscales exceeding .60 indicate measurement overlap, it is perhaps more correct that the magnitude of these correlations is consistent with *a priori* predictions of the relationships between the scales.

4a.4.2 Results

Tables 4.3 and 4.4 present correlations between scales and subscales for the FIM and FIM+FAM. Alpha coefficients are shown in parentheses.

First, consider in Tables 4.3 and 4.4 the triangles of correlations among scales (coloured red). For the FIM and FIM+FAM, correlations between the motor and cognitive scales are below their respective alpha coefficients thus demonstrating unique reliable variance between these scales. This finding provides evidence that the motor and cognitive scales of both instruments are measuring distinct constructs. As predicted, correlations between the motor and total scales and between the cognitive and total scales are substantial due to item overlap. However, correlations between the motor and total scales of the FIM (.97) and FIM+FAM (.93) and between the cognitive and total scales of the FIM+FAM (.91) are very similar to their

respective alphas, indicating that these scales fail the test for unique reliable variance. These findings indicate the lack of an empirical basis for a separate motor scale of the FIM, and separate motor and cognitive scales of the FIM+FAM.

Second, consider the rectangles of correlations between subscales and scales (coloured blue). As predicted, correlations are highest when there is item overlap. Most correlations between subscales and scales are below the alpha coefficients with which they are being compared, indicating the presence of unique reliable variance between scale and subscale. However, some correlations fail the test for unique reliable variance. For the FIM, there is a lack of empirical support for separate self-care, transfer, communication, and social cognition subscales. For the FIM+FAM these findings indicate a lack of empirical support for separate self-care, transfer, psychosocial adjustment, and cognitive function subscales.

Finally, consider the triangles of correlations among subscales (coloured green). All correlations are below the alpha coefficients with which they are being compared, indicating unique reliable variance. However, eight of the 15 correlations among FIM subscales and 11 of the 21 correlations among FIM+FAM subscales exceed the recommended criterion of .60 indicating measurement overlap between subscales. Some of these correlations are particularly high: self-care and transfer for the FIM (.89) and FIM+FAM

(.88), and psychosocial adjustment and cognitive function for the FIM+FAM (.84). These findings indicate little unique reliable variance and, therefore, considerable measurement overlap between these subscales.

4a.5 Are items appropriately grouped into scales and subscales ?

4a.5.1 Method

Two types of analysis, multitrait scaling analysis (302) and principal components analysis (PCA), are undertaken to determine whether the items of the FIM and FIM+FAM are appropriately grouped into scales and subscales.

4a.5.1.1 Multitrait scaling analyses

When instruments such as the FIM and FIM+FAM measure several subconstructs, there should be empirical evidence to show that each item measures one of the subconstructs better than others. Multitrait scaling analyses examine evidence for this on the basis of item convergent and discriminant validity (302). The extent to which each item measures the construct it is hypothesised to measure (item convergent validity) is compared with the extent to which it measures other constructs (item discriminant validity) (246). These analyses follow the logic of the multitrait-

multimethod approach to testing convergent and discriminant validity (303, 304).

Item convergent and discriminant validity for the FIM and FIM+FAM are examined separately for scales and subscales. For each item, correlations with its own scale or subscale (item-own correlation) are compared with correlations with all other scales and subscales (item-other correlations). Item-own correlations determine item convergent validity by indicating the extent to which an item measures the construct it is purported to measure. The difference in magnitude between item-own and item-other correlations determine item discriminant validity by indicating the extent that each item measures other subconstructs.

Item-own to item-other comparisons are interpreted as either definite or probable scaling successes or failures. A definite scaling success, indicating that an item is correctly grouped, is achieved when item-own correlations exceed item-other correlations by more than two standard errors ($1 / \sqrt{n} = > .14$). A definite scaling failure, indicating that items are incorrectly grouped, occurs when item-other correlations exceed item-own correlations by more than two standard errors. A probable scaling success occurs when item-own correlations exceed item-other correlations by two standard errors or less. Similarly, a probable scaling failure occurs when item-other correlations exceed item-own correlations by two standard errors

or less. Results are summarised for each scale and subscale as percent scaling success and failure rates.

Probable scaling successes and failures indicate limited item discriminant validity. They represent items measuring two or more hypothesised constructs to a similar extent that will confound constructs and complicate the interpretation of their scores. Widespread probable scaling successes and failures indicate that hypothesised constructs that are not empirically distinct and suggest that the conceptual model of an instrument needs to be reconsidered. However, probable scaling successes and failures must be interpreted with reference to the actual magnitude of difference between item-own and item-other correlations, sample size, and the number of items in the scale. As sample size determines the standard error of a correlation, substantial differences between item-own and item-other correlations may not reach statistical significance when sample sizes are small. In addition, scaling failures are better tolerated by scales with large numbers of items as there may be enough other items to anchor the construct and to distinguish it from other constructs (302).

4a.5.1.2 Principal components analysis

Exploratory factor analysis of an item pool using the principal components method (PCA) indicates clusters of intercorrelated items within an instrument (components) that are empirically distinct. As components represent separate dimensions of measurement, they are candidates for scales or subscales (305). Empirical support for the measurement model of the FIM and FIM+FAM is provided when components extracted by PCA conform with hypothesised scales and subscales.

Principal components analysis for the FIM and FIM+FAM is undertaken on admission ratings (N = 209). Analyses of FIM data are cross-validated on an independent sample of 367 patients from the NRU audit database. Analyses of FIM+FAM data are cross-validated on two samples generated by randomly dividing the total sample (n = 105 and n = 104). Components extracted by PCA are rotated (varimax) to achieve simple structure. Two standard criteria are used to determine the number of components to rotate: first, all components with eigenvalues greater than 1.0 are retained (306); second, the scree plot of eigenvalues is examined to identify the point at which the negative slope of the curve levels off and begins the scree (307).

Although there are similarities between multitrait scaling analysis and PCA, there are important differences which justify the use of both methods when

examining the measurement models of the FIM and FIM+FAM. Multitrait scaling analysis is an item-level confirmatory analysis that evaluates the appropriateness of *a priori* groups of items being summed to form scales. In contrast, PCA is an item-level exploratory analysis that identifies which items should be summed to from scales. As PCA is not constrained by the conceptual models and assumptions of instrument developers, it helps to identify dimensions of measurement that were not originally hypothesised.

In PCA the trait is defined by the analysis (308). Therefore, artefacts in the data that have little to do with the constructs measured (e.g. sample size) can influence results in ways that distort the interpretation of the underlying constructs (302, 309). Also, if items are added to or removed from an item pool the definitions of the traits extracted by PCA may change (269). In contrast, the scales in multitrait scaling analysis are defined by the investigator and items can be added or removed to examine their relationships with existing scales without altering the underlying constructs (302).

Scales defined in multitrait scaling analysis differ from traits defined by the factors extracted in PCA even if their item content is identical (302). The correlation between an item and a trait defined by PCA (component loading) may not accurately represent the correlation between an item and its scale (item-total correlation).

4a.5.2 Results

4a.5.2.1 Multitrait scaling analysis

Tables 4.5 to 4.8 present the results of multitrait scaling analysis. Tables 4.5 and 4.6 present correlations between items, scales, and subscales for the FIM and FIM+FAM. Tables 4.7 and 4.8 summarise item convergent and discriminant validity as percent definite and probable scaling successes and failures.

First, consider the correlations between items and scales coloured red in Tables 4.5 and 4.6. These correlations determine the extent to which items discriminate between scales. Criteria for item convergent and discriminant validity are satisfied when item-own scale correlations are high and exceed item-other scale correlations by more than two standard errors (> .14). For example, the feeding item of the FIM in Table 4.5 is hypothesised to belong in the motor rather than the cognitive scale. The item-own scale correlation (feeding-motor scale = .70) exceeds the item-other scale correlation (feeding-cognitive scale = .48) by .22. This result is a definite scaling success.

Tables 4.5 and 4.6 indicate good item discriminant validity for FIM scales, but limited item discriminant validity for FIM+FAM scales. All FIM items satisfy criteria for definite scaling successes except grooming, which qualifies as a probable scaling success. In contrast, 21 FIM+FAM items (70%) satisfy criteria for definite scaling successes. Eight items qualify as probably scaling successes and one item (community mobility) qualifies as a probable scaling failure.

Next, consider correlations between items and subscales coloured blue in Tables 4.5 and 4.6. These correlations determine the extent to which items discriminate between subscales. Criteria for convergent and discriminant validity are satisfied when item-own subscale correlations are high and exceed all item-other subscale correlations by more than two standard errors. As an example, consider the dressing lower body item of the FIM in Table 4.5. The item-own subscale correlation (dressing lower body / self-care) is high (.84), supporting its inclusion in this subscale. Four item-other subscale correlations (sphincter care, locomotion, communication, and social cognition) satisfy criteria for definite scaling successes, indicating good discriminant validity between these subscales for the dressing lower body item. However, the dressing lower body-transfer correlation (.87) slightly exceeds the item-own subscale correlation. This result represents a probable scaling failure and indicates that this item measures equally two constructs purporting to be conceptually distinct. Either the dressing lower body item has poor discriminant validity for these two constructs, or, the

constructs measured by the self-care and transfer subscales are not empirically distinct.

Table 4.5 demonstrates limited item discriminant validity for FIM subscales. Only two items, expression and problem solving, fully satisfy criteria for definite scaling successes. Three items (dressing lower body, toileting, and stairs) qualify as probable scaling failures and 15 items register one or more probable scaling successes. Furthermore, the magnitude of differences between item-own and item-other correlations for 62% of these probable scaling successes and failures is small ($\leq .07$) indicating that most FIM items measure two or more constructs equally. The fact that 11 of the 18 FIM items demonstrate limited discriminant validity indicates that some hypothesised constructs are not empirically distinct. For example, all self-care items correlate highly with the transfer subscale suggesting that the self-care and transfer subscales are not measuring empirically distinct constructs. All sphincter care and locomotion items correlate highly with the self-care and transfer subscales suggesting that the sphincter care, locomotion, self-care, and transfer subscales are not measuring empirically distinct constructs. Similarly, three items (comprehension, social interaction, and memory) correlate similarly with the communication and social cognition subscales indicating that these subscales are not measuring empirically distinct constructs.

Table 4.6 indicates that the FIM+FAM has poorer item discriminant validity than the FIM. Only four items (expression, reading, writing, speech intelligibility) fully satisfy the criteria for definite scaling successes. Twenty six-items qualify as having probable scaling successes (bolded), and three of these items (dressing lower body, toileting, and employability) also register probable scaling failures. Eight items register two probable scaling successes, two items (toileting and swallowing) register three probable scaling successes, one item (community mobility) registers four probable scaling successes, and one item (employability) registers five probable scaling errors. A total of 23 item-other subscale correlations lie within one standard error of the item-own subscale correlation for those items, indicating items that measure equally two or more constructs. The fact that 16 items from six of the seven subscales demonstrate limited item discriminant validity suggests that some subscales are not measuring empirically distinct constructs. These subscales are: self-care and transfer; self-care and locomotion; sphincter, self-care and transfer; and psychosocial adjustment and cognitive function.

Tables 4.7 and 4.8 summarise the results of multitrait scaling analyses for the FIM and FIM+FAM as percent definite and probable scaling successes and failures. For the FIM, results indicate good item discriminant validity for scales but limited item discriminant validity for subscales. For the FIM+FAM, results indicate limited item discriminant validity for scales and subscales. These findings provide strong empirical support for grouping

FIM items into motor and cognitive scales, but weak support for grouping FIM+FAM items into scales and for grouping items of both instruments into subscales.

4a.5.2.2 Principal components analysis

Table 4.9 presents results from two principal components analyses of FIM admission scores. Results from the two analyses are different as the first analysis extracts two components and the second analysis four components.

The first analysis (PCA-1) extracts two components with eigenvalues > 1.0 which satisfy the scree test. These components account for 71.6% of the total variance. The first component accounts for 59.4% of the total variance with 13 items loading on this component. The second component accounts for 12.2% of the total variance with five items loading on this component. The items constituting both components are identical to the motor and cognitive scales of the FIM. These results support the grouping of FIM items into motor and cognitive scales but do not support FIM subscales as separate dimensions of measurement.

The second analysis (PCA-2) extracts four components with eigenvalues > 1.0 that satisfy the scree test. These account for 74.6% of the total

variance. Eight items from the self-care, transfer, and locomotion subscales load on component 1. Four items from the self-care subscale load on component 2. All five items of the cognitive scale load on component 3. The two items from the sphincter subscale load on component 3. The bathing item loads onto components 1 and 2 equally indicating that this item has limited discriminant validity.

The results of PCA-2 do not provide strong support for the measurement model of the FIM. Although results support the cognitive scale and sphincter subscale as separate dimensions of measurement, they do not support the motor scale or five other subscales (self-care, transfer, locomotion, communication, or social cognition) as separate dimensions of measurement. Results suggest that items in the motor scale might better be grouped in three subscales: lower limb function; upper limb function; and sphincter care. As bathing involves both upper and lower limbs the finding that is loads equally on components 1 and 2 is intuitively sound.

Table 4.10 presents results from three principal components analyses (total sample and random split-halves) of FIM+FAM admission scores. The results of the three analyses are very similar. All three PCAs extract four components with eigenvalues > 1.0 which satisfy the scree test. These four components account for approximately 77% (range 75.8% to 79.5%) of the total variance. The item composition of the components extracted by the

three PCAs is consistent for all four components. Most items from the hypothesised self-care, transfer, and locomotion subscales load on component 1. All items from the psychosocial adjustment and cognitive function subscales load on component 2. Items from the communication subscale load on component 3. Three items (bladder management, bowel management, and swallowing) load on component 4. Five items (feeding, grooming, community mobility, comprehension, employability) consistently load on two or more components indicating that they have poor discriminant validity.

Although subscale items usually load on the same component, the items comprising these components are not consistent with the scale or subscale structure of the FIM+FAM. Results suggest that the items of the motor scale comprise two dimensions of measurement rather than either a single scale or four distinct subscales as hypothesised. Similarly, items of the cognitive scale appear to comprise two dimensions of measurement rather than a single scale or three distinct subscales as hypothesised.

4a.6 Summary of results

4a.6.1 FIM

Results from item analyses provide evidence that all FIM scales and subscales represent internally consistent groups of items which can be legitimately summed without weighting to generate scores. However, there is evidence of item redundancy in all three scales and in four of the six subscales. Empirical evidence indicates that the 13-item motor and 5-item cognitive scales measure distinct subconstructs and that items in these scales adequately discriminate between the subconstructs. However, the motor scale measures a similar construct to the total scale. Finally, results provide evidence that the FIM subscales do not represent distinct dimensions of measurement, that there is limited item discriminant validity, and that a different grouping of items into subscales might be more appropriate.

4a.6.2 FIM+FAM

Results from item analyses provide evidence that all FIM+FAM scales and subscales represent internally consistent groups of items which can be legitimately summed without weighting to generate scores. However, there is evidence of item redundancy in all three scales and in six of the seven subscales. Empirical evidence indicates that the 16-item motor and 14-item

cognitive scales measure distinct subconstructs, but that nine items (6 motor, 3 cognitive) do not adequately discriminate between the subconstructs. In addition, both the motor and cognitive scales measure a similar construct to the total scale. Finally, results provide evidence that the FIM+FAM subscales do not represent distinct dimensions of measurement, that there is limited item discriminant validity, and that a different grouping of items into subscales might be more appropriate.

4b Development and psychometric evaluation of a short-form version of the FIM

The second question addressed in this chapter is the feasibility of developing a short-form measure. Results of the psychometric analyses of the FIM and FIM+FAM and comparison with the Barthel Index presented in Chapter 3 raised the question of item redundancy: item analyses confirm this is in both measures. As the FIM and FIM+FAM have been shown to be highly similar instruments which can be considered alternate form measures of the same construct, only FIM data are used for the analyses reported in this section.

Based on the results of psychometric analyses presented earlier in this chapter, item reduction analyses are undertaken to develop a short-form

version of the FIM. Items are then grouped into scales, followed by tests of scaling assumptions and the evaluation of psychometric properties.

4b.1 Item reduction

4b.1.1 Method

Items are eliminated if they fail to satisfy previously defined criteria for acceptability and reliability or are redundant. Items are not acceptable if responses are poorly distributed among item response categories or if maximum endorsement frequencies, floor, or ceiling effects exceed 80%.

Items fail to satisfy criteria for reliability if corrected item-total correlations are <.30 (217), or if intra- or inter-rater reproducibility is <.70. Items are defined as redundant if inter-item correlations exceed .70 (297). The decision as to which of the two items to delete depends upon a comparison of descriptive statistics, reliability, responsiveness, and clinical importance. As the FIM is an evaluative instrument, item responsiveness is the primary criterion for item selection.

4b.1.2 Results

Table 4.11 presents descriptive statistics, reliability, and responsiveness for FIM items. All items satisfy the criteria discussed above. Table 4.12 presents the 31 item-intercorrelations that exceed .70. Fifteen different items are involved, the remaining three items, bladder management, bowel management, and walking have intercorrelations with all other FIM items that are < .70.

No items were deleted due to poor discrimination between subjects or poor reliability. Therefore, for each pair of items with intercorrelations $\geq$ .70, the item with the best responsiveness is retained and the other deleted. Ten items were removed due to item redundancy, leaving eight items to form the short-form FIM-8: feeding, bladder management, bowel management, shower transfer, stairs, walking, social interaction, and memory.

4b.2 Development of scales

4b.2.1 Method

Two methods were used to form scales after item reduction: conceptually and empirically derived scales. First, items in the FIM-8 were grouped on the basis of a conceptual knowledge of disability. Next, item-level

exploratory factor analysis was performed using the principal components methods. Components with eigenvalues > 1.0 that satisfied the scree test were varimax rotated to achieve a simple structure. Components indicate clusters of intercorrelated items that are empirically distinct and, therefore, candidates for separate scales. Principal components analysis was performed on the whole study sample (N = 209) and cross-validated on two samples generated by randomly dividing this sample (n = 105, n = 104).

4b.2.2 Results

4b.2.2.1 Conceptually derived scales

The conceptual model of disability generated three methods of grouping FIM-8 items into scales. First, an overall scale comprising all eight items was formed to measure global disability. Second, items were grouped into two scales: a 6-item motor scale (feeding, bladder management, bowel management, shower transfer, stairs, and walking), and a 2-item cognitive scale (social interaction, and memory). Third, items were grouped into three scales: a 4-item physical scale (feeding, shower transfer, stairs, and walking), a 2-item sphincter scale (bladder management, bowel management), and a 2-item cognitive scale (social interaction, and memory). This conceptual grouping is based on results from the PCA reported in the previous section which suggested sphincter function is an independent dimension.

4b.2.2.2 Empirically derived scales

Table 4.13 presents results from three principal components analyses of FIM-8 admission scores. The first analysis (PCA-1) extracts two components with eigenvalues > 1.0 which satisfy the scree test and which account for 67.9% of the total variance. The first component, which consists of four items measuring physical function, accounts for 52.0% of the variance. The second component , which consists of four items measuring sphincter and cognitive function, accounts for 15.9% of the variance.

Cross-validation largely supports the findings of PCA-1. In PCA-2 two components with identical item content are extracted. In PCA-3 two components are also extracted but three items (feeding, bladder management, and bowel management), load equally and substantially onto both components. The other five items load on the two components in the same manner to PCA-1 and PCA-2.

4b.3 Evaluation of summated rating scales

4b.3.1 Method

Item groups defined on conceptual and empirical grounds were evaluated in terms of their appropriateness as simple summated rating scales. First, the internal consistency of scales was examined to determine if all items were indeed measuring the same construct and if items can be summed without weighting to generate scores. Second, intercorrelations between scales were examine to determine the extent to which scales measure different constructs. Third, multitrait scaling analyses were performed to determine scaling success rates. The optimal scaling method is defined as method which achieves the highest internal consistency, largest unique reliable variance between scales, and the highest scaling success rate.

4b.3.2 Results

4b.3.2.1 Internal consistency

Table 4.14 presents internal consistency estimates for the four methods of scaling FIM-8 items. All scales have high internal consistency, with alpha coefficients exceeding the minimum requirement of .70 for group comparisons. These results indicate that items in each scale measure the

same underlying construct. In addition, item-total correlations for all scales are high indicating that Likert's criteria of equivalent item means and variances can be considered satisfied. These results indicate that there is no clear advantage of any one method of scaling FIM-8 items.

4b.3.2.2 Intercorrelations between scales

Table 4.15 presents intercorrelations between scales (with alpha coefficients in parentheses) for the four methods of scaling FIM-8 items. As expected all scales correlate highly with the FIM-8 total scale due to item overlap. Inter-correlations between scales derived by methods 2, 3, and 4 are all well below the alpha coefficients for these scales indicating unique reliable variance for each scale. These results indicate that the scales derived by the four methods measure related but different constructs, and that there is no clear advantage of any one method.

4b.3.2.3 Multitrait scaling analyses

Table 4.16 summarises multitrait scaling analysis for the three methods of grouping FIM-8 items with two or more scales. Method 2, the grouping of FIM-8 items into a 6-item motor scale and a 2-item cognitive scale, has the highest proportion of definite scaling success rates. These findings indicate that this method is superior to the others in terms of item convergent and

discriminant validity. Methods 3 and 4 have limited item convergent and discriminant validity. These findings provide only limited support for the integrity of scales developed using methods 3 and 4.

4b.4 Psychometric evaluation of the FIM-8

4b.4.1 Method

The short-form FIM-8 (total, 6-item motor, and 2-item cognitive scales) was evaluated for acceptability, reliability, validity, and responsiveness. Acceptability was evaluated by examining scale score distributions. Three types of reliability were estimated: internal consistency, intra-rater, and inter-rater reproducibility. Two types of validity, concurrent and construct validity, were examined. Concurrent validity was assessed by examining product-moment correlations between FIM-8 scales, the original 18-item FIM, and the FIM+FAM. Internal construct validity was determined by examining internal consistency and intercorrelations between scales for the FIM-8. External construct validity was determined by examining product-moment correlations between the FIM-8 and other measures. Convergent and discriminant validity were determined by examining the magnitude and pattern of correlations with measures of disability, handicap, health status, psychological well-being, and neuropsychological functioning. Discriminant validity was also evaluated by examining correlations between the FIM-8, age, and sex. Responsiveness was determined by calculating an effect

size and comparing this to the original 18-item FIM and the FIM+FAM to determine relative responsiveness.

4b.4.2 Results

4b.4.2.1 Acceptability

Table 4.17 presents score distributions for FIM-8 scales. These results indicate that FIM-8 scales adequately represent the range of severity of disabilities in patients in this sample and demonstrate good variability. However, the cognitive scale has a notable ceiling effect of 27.8% which is above the recommended maximum of 15%. These findings indicate that the total and motor scales of the FIM-8 are more acceptable than the cognitive scale.

4b.4.2.2 Reliability

Table 4.14 and 4.18 present reliability estimates for the FIM-8 scales. Table 4.14 indicates that minimum criteria for internal consistency are satisfied. Table 4.18 indicates that criteria for reproducibility are satisfied. These findings provide evidence for the reliability of FIM-8 scales.

4b.4.2.3 Validity

Concurrent validity. Table 4.19 presents correlations between the FIM-8, FIM, and FIM+FAM. Correlations between the total, motor, and cognitive scales of the FIM-8 and corresponding FIM and FIM+FAM scales are high. These findings indicate that the FIM-8 can be considered an alternate form measure for the FIM and FIM+FAM.

Internal construct validity. Internal consistency estimates presented in Table 4.14 provide evidence for internal construct validity of the FIM-8. Table 4.20 presents intercorrelations (with alpha coefficients in parentheses) for the three FIM-8. As discussed previously, the correlation between the motor and cognitive scale is substantially lower than the alphas for these scales, demonstrating the presence of unique reliable variance. These results provide evidence for internal construct validity by indicating that the motor and cognitive scales measure different but related constructs.

External construct validity. Table 4.19 presents correlations between the FIM-8, FIM, and FIM+FAM. All correlations are high, thus providing evidence of convergent validity for FIM-8 scales. In addition, the pattern of correlations in Table 4.19 provides evidence of discriminant construct validity for the FIM-8 scales. For example, of the three FIM-8 scales, the

highest correlation with the FIM total scale is for the FIM-8 total scale. This pattern is present for all FIM scales against both the FIM and FIM+FAM.

Table 4.21 presents correlations between FIM-8 scales and other measures of disability, handicap, health status, psychological distress, and age. The magnitude and pattern of these correlations provides evidence supporting the convergent and discriminant validity of the total, motor, and cognitive scales of the FIM-8. For example, correlations are highest with the four disability scales (BI, MBI, EDSS, OPCS) and lowest with mental health and psychological distress (SF-36 MCS and GHQ). In addition, correlations between the FIM-8 and the disability scales are higher for the motor than the cognitive scale.

Table 4.22 presents correlations between the FIM-8 and neuropsychological measures. For four neuropsychological measures (MMSE, WAIS-VIQ, WCST, VESPAR), correlations are highest with the FIM-8 cognitive scale and lowest with the FIM-8 motor scale. These findings provide evidence of convergent and discriminant validity for the FIM-8 cognitive scale. However, the remaining three neuropsychological measures (Halstead Booklet Category Test, California Verbal Learning Test, Visual Recognition Memory Test) do not demonstrate this pattern of correlation with FIM-8 scales.

Table 4.23 presents correlations between the FIM-8, age, and sex. These are low indicating that FIM-8 scores are not biased by age and sex, and providing further evidence of discriminant validity.

4b.4.2.4 Responsiveness

Table 4.24 presents results for responsiveness of the FIM-8. Change scores for all scales are negative, indicating an improvement in disability at discharge. Effect sizes indicate that the motor and total scales show medium responsiveness whilst the cognitive scale shows poor responsiveness. The responsiveness of the FIM-8 is equivalent to that of the FIM and FIM+FAM.

4b.5 Summary of results

Results support the appropriateness and psychometric adequacy of a short-form version of the FIM, the FIM-8. The most valid grouping of the 8 items is into three scales: a total scale, a 6-item motor scale, and a 2-item cognitive scale. All three scales are psychometrically equivalent to comparable FIM scales.

Chapter 5

Comparison of Methods of Evaluating Responsiveness: Method and Results

Analyses reported in Chapters 3 and 4 examined the responsiveness of FIM and FIM+FAM scales. However, there is no consensus as to which of the available statistical methods for reporting responsiveness should be used. Consequently, different studies use different statistical methods for evaluating responsiveness, thus making comparisons difficult. Furthermore, there is little information about the clinical implications of using different methods for reporting responsiveness.

The objective of the analyses presented in this chapter is to compare five different methods of evaluating responsiveness. The methods are compared in two ways: first, how each method rank-orders the six scales of the FIM and FIM+FAM, and second, the relative responsiveness each scale compared with an arbitrary standard.

5.1 Method

Responsiveness was evaluated using five statistical methods. Data from the 194 patients in this study who completed the inpatient neurorehabilitation programme are used for these analyses. As is previous studies (312, 318), results for each statistical method are reported as the magnitude of instrument responsiveness and the rank ordering of the six FIM and FIM+FAM scales. For all methods, higher values indicate greater responsiveness and rank ordering is from 1 (most responsive) to 6 (least responsive).

The five methods of evaluating responsiveness are also compared using an approach not previously adopted. This approach examines the relative responsiveness, in proportional terms, of the six instruments for each statistical method. It is calculated by dividing the value for each scale by the value of a nominated arbitrary standard, defined here as the FIM total scale. Relative ratios of 1.0 indicate instruments that are as responsive as the standard, whilst values exceeding 1.0 indicate better responsiveness compared with the standard and values less than 1.0 indicate poorer responsiveness compared with the standard. This approach to comparing methods of evaluating responsiveness complements rank ordering by indicating the extent to which instruments differ in their responsiveness.

Five methods are used to evaluate responsiveness:

***t*-statistics** generated from paired samples *t*-tests on admission and discharge scores for each subject are computed as the mean change score divided by the standard error of change scores (SD change ÷ $\sqrt{n}$) (310). The variation in change scores is the reference against which the magnitude of change is judged. There are two limitations of *t*-statistics as indices of responsiveness. First, they contain an adjustment for sample size and can be misleading when sample sizes vary. Second, they fail to account for variation in scores for clinically stable subjects (310).

Relative efficiency is the extent to which one scale is more or less efficient at detecting change in disability over time relative to another scale (133). Squared *t*-statistics are computed from paired samples *t*-tests. One scale is nominated as the arbitrary standard, in this study the FIM total scale. The relative efficiency of each scale is then computed by dividing its squared *t*-statistic by the squared *t*-statistic for the FIM total scale. Values of 1.0 indicate scales that are as efficient at detecting change over time as the FIM total scale, whilst values greater (or less) than 1.0 indicate scales that are more (or less) responsive than the FIM total scale. By relating each scale to a standard, relative efficiency calculations offer the advantage of indicating instrument responsiveness in proportional terms. However,

relative efficiencies have the same limitations as *t* -statistics and can only be used when comparing instruments in the same dataset.

Effect size, as defined by Kazis *et al.* (132), is computed as the mean change score for each scale divided by the standard deviation of admission scores. The variation of baseline score is the reference against which the magnitude of change is judged. Kazis *et al.* recommend that the clinical relevance of effect sizes can be interpreted using the arbitrary criteria originally proposed by Cohen (.20 = small; .50 = medium; .80 = large; 132, 311). The limitation of effect sizes as indices of responsiveness is that they fail to account for variation in change scores and in scores of stable subjects over time.

Standardised response mean is computed as the mean change score divided by the standard deviation of change scores (135). Variation of change is the reference against which the magnitude of change is judged. This has direct implications for sample size determination as the ratio of sizes required to detect a given clinical effect is equal to the square of the ratio of standardised response means (292). Liang (135) recommends that Cohen's criteria, defined above, can be used to interpret the clinical importance of standardised response means. The limitation of the standardised response mean as an index of responsiveness is that it does not account for variation in scores for clinically stable subjects.

Guyatt *et al.'s* responsiveness index is calculated as the mean change score for respondents who changed divided by the standard deviation of change scores for respondents who did not change (110). Variability in clinically stable subjects is the reference against which the magnitude of change is judged (292). Consistent with the approach taken in other studies (312), Guyatt *et al.'s* responsiveness index is calculated as the mean change score in all respondents (not just the subsample who changed), divided by the standard deviation of change scores for respondents in the test-retest sample.

There are a number of limitations associated with the responsiveness index as a method determining responsiveness. First, basing the numerator and denominator on different samples is subject to bias as it assumes similar variance in the two samples (134). Second, this method does not take into account systematic change that may occur in patients whose health status is stable (291). Third, it does not consider variability in the change group (136). Finally, within-patient variability on a measure in patients whose health status is stable (the denominator in Guyatt *et al.'s* method) increases as the length of time between assessments increases (291).

Guyatt *et al.* suggest that the most appropriate index of responsiveness is computed by dividing the minimum change score considered clinically important (minimum clinically important difference) by the variability in scores of stable subjects (110). However, they acknowledge that for many new and established instruments the minimum clinically important difference is unknown. They recommend that an initial estimate of responsiveness can be determined by comparing the within-person standard deviation to the change score observed after an intervention of known efficacy (110). This is the standardised response mean.

5.2 Results

Table 5.1 presents results of responsiveness analyses using the five statistical methods described above. The numerical estimates generated by the various methods differ in magnitude. This is expected as each method of calculating responsiveness, except relative efficiency, relates the same mean change score to a different denominator. That effect sizes and standardised response means generate different numerical estimates is notable as Cohen's criteria have been proposed for the interpretation of both methods. Therefore, applying Cohen's criteria can lead to different conclusions about the responsiveness of an instrument. For example, the responsiveness of the FIM total score is moderate as assessed by the effect size and large when calculated using the standardised response mean.

For all scales, responsiveness assessed in terms of standardised response means is higher than when calculated on the basis of effect sizes. The finding indicates that the standard deviation of change scores is consistently less than the standard deviation of admission scores. Liang suggests that this finding indicates that the correlation between admission and discharge scores is high (292). When correlations between admission and discharge scores are low (<.50) effect sizes exceed standardised response means, whereas effect sizes and standardised response means are equal when this correlation is .50.

Table 5.2 presents a rank ordering of the responsiveness of FIM and FIM+FAM scales for each method (1 = most responsive). In addition, the relative responsiveness of each scale compared to the FIM total scale is also presented. Rank ordering produces similar results across all methods. The only difference is that three methods (*t* -statistic, relative efficiency, standardised response mean) show the FIM motor scale to be more responsive than the FIM+FAM motor scale. For the other two methods this order is reversed. As these two instruments have almost identical responsiveness by all five statistical methods, this is of little clinical relevance. Relative responsiveness produces more variability between the different methods of evaluating responsiveness. These results indicate that for the five methods of reporting responsiveness studied the choice of

statistic has little influence on instrument responsiveness when considered in terms of rank ordering.

5.3 Summary of results

The five methods of reporting responsiveness produce numerical estimates that vary in magnitude. Applying Cohen's criteria to effect sizes and standardised response means leads to different clinical interpretations of the responsiveness of instruments. The five methods produce very similar rank ordering of FIM and FIM+FAM scales, their relative ratios are similar though more variable.

Chapter 6

Discussion

The role of the NHS is to deliver modern effective health care to people in the United Kingdom within a limited budget. To achieve this aim, new and existing therapeutic interventions must be evaluated and compared in the environments in which they are used (313). This requires the use of health outcome measures that combine scientific soundness and clinical usefulness. The need for rigorous outcomes measurement cannot be overstated: information collected using health measures influences decisions that affect patient welfare and guide major expenditure of public funds (127).

Neurorehabilitation is a complex resource-consuming intervention with a high service demand and a limited scientific basis (149). Careful evaluation of neurorehabilitation is needed urgently to determine its effectiveness relative to other interventions and to underpin future research and development. While level of disability is the primary outcome for evaluating rehabilitation, disability measurement is poorly developed (84). Among numerous existing measures, the FIM and FIM+FAM have achieved particular importance and widespread use. Although the clinical usefulness of these measures is agreed, their scientific properties have not been fully

documented. The objective of this study was to evaluate comprehensively the psychometric properties of these two measures. Given the importance of evaluating the responsiveness of health outcome measures, this topic was given particular attention.

This study addressed three questions. First, are the FIM and FIM+FAM rigorous measures of disability in neurorehabilitation? This question included a comparison of the measurement properties of both instruments in stroke and MS and with the Barthel Index. Second, is there empirical support for the conceptual models on which both measures are based? This question included an evaluation of the feasibility of developing a psychometrically sound short-form measure. Third, how do five methods of evaluating responsiveness compare?

Results show that the FIM and FIM+FAM are acceptable, reliable, valid, and responsive measures of disability in neurorehabilitation. Furthermore, they have similar measurement properties in stroke and MS patients. However, neither instrument shows any psychometric advantage as an evaluative measure over the Barthel Index. Moreover, both the FIM and FIM+FAM show item redundancy, limited item discriminant validity, and inadequate support for the hypothesised subscales. An 8-item short-form FIM was developed that shows similar psychometric performance to the 18-item FIM and 30-item FIM+FAM. The five methods of evaluating

responsiveness rank order scales similarly, but generate numerical estimates of different magnitude.

This study makes four major contributions to health measurement. First, findings demonstrate that the FIM and FIM+FAM (and also the Barthel Index) are suitable for measuring disability in clinical practice and research, and indicate which scores should be reported. Second, recommendations are made for more rigorous standards for instrument development and evaluation before introduction into practice. Third, results suggest that conceptual models of disability need to be refined. Finally, the study emphasises the need for consensus in evaluating responsiveness and makes recommendations for reporting responsiveness.

This chapter summarises study results and compares them with findings from previous studies. The limitations of the study and implications for clinical practice and research are examined. Finally, future research directions are suggested.

6.1 Summary of results

6.1.1 Psychometric evaluation of the FIM and FIM+FAM

This study provides clear evidence that the FIM and the FIM+FAM are acceptable, reliable, valid, and responsive measures of disability in patients undergoing inpatient neurological rehabilitation. However, the total and motor scales are more responsive than the cognitive scale for both instruments. These findings indicate that the FIM and FIM+FAM are suitable for use as disability outcome measures in audit, clinical practice, and research in groups similar to the patients studied.

Results also demonstrate that the FIM and FIM+FAM are psychometrically sound measures of disability in stroke and MS patients. There are, however, some differences between the performance of the two instruments. The FIM cognitive scale is less acceptable in MS than in stroke patients due to high ceiling effects. All FIM and FIM+FAM scales are more responsive in stroke than in MS patients, suggesting that responsiveness may be disease dependent. Correlations with handicap are higher for MS than for stroke patients indicating better discriminant validity in stroke patients.

The FIM, FIM+FAM, and Barthel Index have similar measurement properties in patients undergoing inpatient neurorehabilitation. Corresponding scales of the FIM and FIM+FAM have equivalent psychometric properties, indicating that the addition of FAM items to the FIM does not improve disability measurement. Moreover, the Barthel Index has equivalent psychometric properties to the motor scales of the FIM and FIM+FAM, except for measurement precision, indicating that the presumed advantages of the newer FIM and FIM+FAM are not supported by empirical evidence. Furthermore, the similarity in psychometric properties between the Barthel Index and the total and motor scales of the FIM and FIM+FAM calls into question the claim that the Barthel Index is too crude, simple, and insensitive (unresponsive) to be used to evaluate clinical practice or in research (163, 165, 166). In this study, the Barthel Index demonstrates good acceptability, internal consistency reliability, construct validity, and responsiveness: there is no evidence that the FIM or FIM+FAM perform better from a psychometric point of view.

This is the first study to comprehensively evaluate the full range of psychometric properties of the FIM and FIM+FAM in the same sample, to compare these properties in stroke and MS, and to compare them head-to-head with the Barthel Index. In addition, this study is the most comprehensive psychometric examination of the Barthel Index to date. Results from this study are similar to those reported in previous studies examining aspects of the psychometric properties of the FIM and FIM+FAM.

For the FIM, high internal consistency and inter-rater reproducibility have been demonstrated for all three FIM scales (195, 201-203, 205, 206). For both the FIM and FIM+FAM, high correlations between the total scales and the OPCS disability scale (216), and similar intercorrelations between the motor and cognitive scales (191) have been reported.

Findings from this study are consistent with previous studies that have evaluated aspects of the measurement properties of the FIM in stroke and in MS patients. For stroke patients, the FIM total scale has been shown to be acceptable (205, 314, 315), and all three FIM scales have been shown to be internally consistent (205). For MS patients, the FIM total scale has been shown to be acceptable (197, 201, 208), internally consistent (201), reproducible (201), and highly correlated with the EDSS (197, 201). There are no studies which have compared the performance of the FIM+FAM in stroke and in MS patients.

Although this is the first study to report a comprehensive head-to-head comparison of the measurement properties of the FIM, FIM+FAM, and Barthel Index, previous studies report similarities between these three instruments. High correlations have been shown between the total scales of the FIM and FIM+FAM (216), and between the motor scales of the FIM and FIM+FAM (191). These convergent validity estimates are the same as those found in this study, confirming the similarity of these scales.

There are, however, some differences between results of this study and previous research. For example, Hall *et al.* (191) report a lower correlation between the cognitive scales of the FIM and FIM+FAM (.84 versus .97). Similarly, McPherson and Pentland (216) report lower correlations between the Barthel Index and the FIM total score (.64 versus .95), and between the Barthel Index and the FIM+FAM total score (.53 versus .90). These findings from previous studies suggest that the FIM, FIM+FAM, and Barthel Index are measuring related but distinct health constructs. However, differences in the score distributions of the instruments, which attenuate correlations between instruments (316), may explain why these correlations are lower than those obtained in this study. In the Hall *et al.* study, the distribution of FIM+FAM cognitive scores is strongly positively skewed, whereas FIM cognitive scores more closely approximate a normal distribution. In the McPherson and Pentland study, Barthel Index scores have a ceiling effect of 69% and are, therefore, strongly negatively skewed. Conversely, FIM and FIM+FAM scores more closely approximate a normal distribution.

6.1.2 Evaluating conceptual models using item analysis

Item analysis of the FIM and FIM+FAM demonstrates item redundancy, subscale overlap, limited item discriminant validity, and dimensions of measurement that are inconsistent with hypothesised scales. Thus, the

empirical evidence does not fully support underlying conceptual models. Item redundancy also suggests that short-form versions of both instruments can be developed. On the basis of further psychometric analyses, ten redundant items were eliminated from the FIM to produce an 8-item short-form version. Analyses indicate that these items are best grouped into three scales (an 8-item total scale, a 6-item motor scale, and a 2-item cognitive scale) that are psychometrically equivalent to corresponding FIM and FIM+FAM scales.

No previous studies have examined the conceptual models of the FIM or FIM+FAM by performing comprehensive item analyses. In fact, no previous studies have had the specific aim of examining the conceptual models of these instruments. However, within the context of different objectives, other investigators have undertaken limited item analyses for the FIM and the results of all-bar-one of these studies tend to support its conceptual model. The internal consistency of the three FIM scales is supported by high alpha coefficients (195, 201, 205) and high item-total correlations (205). Support for the FIM motor and cognitive scales as separate dimensions of measurement is provided by intercorrelations between these two scales (191), principal components analysis specifying a two-component solution (205), multitrait scaling tests (205), and the analysis of Rasch-transformed FIM scores (317).

The results of one study call into question the conceptual model of the FIM. Recently, Stineman *et al.* (206) performed item-level exploratory factor analysis on FIM scores for twenty impairment groups including seven neurological impairments. In contrast to their previous study based on the same dataset (205), Stineman *et al.* (206) did not specify the number of components to be extracted by the principal components analysis. None of the solutions extracted, which have between two and four components, support the scale and subscale structure of the FIM. Rather than question the conceptual model of the FIM, the authors of this study suggest that different combinations of items may need to be used in different impairment groups.

The findings of the present study do not support the conceptual model of the FIM because crucial analyses have been undertaken that have not been reported before. For example, no previous study of the FIM has reported item intercorrelations, intercorrelations between subscales, or item convergent and discriminant validity for subscales. It is the results of these three analyses, as well as those of a principal components analysis when the number of components to extract are not specified, that cast doubt on the conceptual model of the FIM.

The feasibility of developing short-form versions of the FIM or FIM+FAM has not been investigated previously. However, results from other studies

suggest item redundancy in the FIM. It has been suggested that alpha coefficients exceeding .90 to .95 may be indicative of item redundancy (286, 298), particularly for scales having fewer than ten items (98). The consistent finding in other studies that alpha coefficients for FIM scales exceed .90 (195, 201, 205) suggests the possibility of item redundancy and the feasibility of developing short-form measures.

6.1.3 Comparison of methods of evaluating responsiveness

Five methods of reporting responsiveness used in this study produce different results in terms of the absolute value of their numerical estimates. However, all five methods produce similar rank orderings of the six scales of the FIM and FIM+FAM. In fact, these five methods of reporting responsiveness generate only two rank orderings of the scales: three methods (*t* -statistics, relative efficiency, standardised response mean) produce one rank ordering, and two methods (effect size, Guyatt *et al.'s* responsiveness index) the other rank ordering. The difference between these two rank orderings is simply a juxtaposition of two scales that have very similar responsiveness. That is, the FIM+FAM motor scale is marginally more responsive than the FIM motor scale when determined using three methods, but marginally less responsive than the FIM motor scale when responsiveness is determined using the other two methods. These findings suggest that the choice of method of evaluating responsiveness has little consequence when the aim of a head-to-head

comparison of instruments is simply to determine which instrument is most responsive. However, as the relative ratios of responsiveness produce more variability between methods, the choice of statistic does indeed have an influence over the relative responsiveness of instruments. This may be important for sample size calculations.

Two previous studies (312, 318) which compared different methods of evaluating responsiveness produced conflicting results. Three statistical methods used in both studies (effect size, standardised response mean, Guyatt *et al.'s* responsiveness index) are also used in the present study. Stucki *et al.* (318) demonstrated almost identical rank ordering of four health measures using five indices of responsiveness. They concluded that the choice of statistic was of little consequence. In contrast, Wright and Young (312) demonstrate different rank orderings for the responsiveness of fourteen health measures using five indices of responsiveness. They conclude that the choice of statistic influences the relative responsiveness of measures. A close examination of the results from these two studies provides a possible explanation for their different conclusions. In the Stucki *et al.* study, the relative responsiveness of instruments differs much more than in the Wright and Young study. Consequently, the rank ordering of instruments in the Wright and Young study is influenced by small differences in the relative magnitude of the numerical estimates produced by different methods. A similar finding to the FIM total and FIM+FAM total scales in this study.

As rank ordering does not quantify relative responsiveness, this method of comparing instruments can be misleading. For example, in the Wright and Young study the responsiveness of the SF-36 Physical Functioning scale is rank ordered four places above the Western Ontario and McMaster Scale. However, the standardised response means are 1.1 and .90, respectively, indicating similar responsiveness. In contrast to rank ordering, relative ratios quantify the relative responsiveness of instruments and provide valuable information for investigators choosing between instruments.

6.2 Study limitations

The results of this study, which are based on a heterogeneous sample of neurologically disabled patients from three centres in and around London, may not be generalisable to all patients undergoing neurological rehabilitation in the UK. Similarly, results may not be generalisable to all stroke and MS patients who are undergoing inpatient rehabilitation. Rehabilitation units throughout the UK differ markedly in the type of rehabilitation programmes offered, the casemix and level of disability of patients, staffing levels and expertise, and facilities. Furthermore, patients are usually selected on the basis of their suitability for a specific rehabilitation programme. Although this study was conducted at three independent units with different expertise, the patient sample is almost

entirely from the south of England. In generalising the results from this study to other local samples, consideration should be given to differences in level of disability of patients.

The disability level of patients in this study is within the range of disability reported by other investigators. Table 6.1 shows that patients in this study are more disabled than two studies of mixed patients in Scotland (214, 216), and less disabled than two studies of mixed patients in the US (200, 205). The disability level of stroke patients is similar to one other study (314), lower than patients in another study (205), and higher than patients in another study (211). The disability level of MS patients is similar to that reported in other studies (197, 201, 208).

The method used to derive FIM scores in this study may be a potential limitation. As the 18 items of the FIM are contained within the 30 items of the FIM+FAM, it is standard practice to rate all 30 items together and derive FIM and FIM+FAM scores subsequently (191, 214, 216, 319). This was the method used in this study. Alternatively, the FIM can be administered and scored independently. As different methods of administering an instrument should be evaluated independently (130), a preliminary investigation of the psychometric properties of the two methods of administering the FIM (stand alone versus as part of the FIM+FAM) was performed. FIM scores from the 118 patients in this study from the NRU were compared with all patients in

the NRU audit database (n = 728) in terms of alpha coefficients (total, scale), intercorrelations among and between scales, item intercorrelations, and correlations with other variables (Barthel Index, EDSS, age). Results (Table 6.2) indicate that the psychometric properties of the FIM do not differ between the two methods of administration.

Another potential limitation concerns the method of administration of the FIM, FIM+FAM, and Barthel Index. As patients were rated by therapists who were providing treatment, staff ratings may have been biased by the need to demonstrate improvements in disability following rehabilitation. Consequently, it is possible that patients may be rated lower (more disabled) on admission and higher on discharge. This would have the effect of attenuating correlations with other measures (88), and overestimating the responsiveness of the FIM and FIM+FAM by increasing the magnitude of the change scores. However, bias due to raters is likely to be minimal for two reasons. First, the aim of this study was not to examine the effectiveness of rehabilitation. Second, the composition of the treatment teams was highly variable and staff turnover moderate.

The method used to evaluate reproducibility may have overestimated FIM and FIM+FAM reliability. Given the short intra-rater reproducibility interval, reliability may have been overestimated due to therapists recalling their initial ratings. However, as each therapist rated many patients and both

instruments have a large number of items, it is unlikely that observers were able to recall their numerous ratings with any degree of accuracy. Similarly, the method used to assess inter-rater reproducibility may also have overestimated reliability. As both the team consensus and study co-ordinator ratings were generated from the same observers (the latter based on interviews of all members of the treating team), the paired ratings were not entirely independent. Nevertheless, the two methods of rating the same patient were different.

Agreement between raters from the three clinical sites, i.e. the inter-site reproducibility of FIM and FIM+FAM scores, was not examined in this study. As both instruments are used widely in different settings where raters are often trained locally, inter-site variability is a potential source of random error affecting FIM and FIM+FAM scores. Unfortunately, there are few clinical settings in which this aspect of reproducibility can be studied easily as it requires patients to be rated by independent teams at two or more sites. However, a previous study (200) has addressed this problem and reports results that are adequate for group comparisons.

The external construct validity of the FIM and FIM+FAM was examined in subsamples of the total sample which varied in size and casemix. For example, tests of neuropsychological functioning were administered only at the NRU which consists predominantly of MS patients. Ideally, all

instruments used to examine the validity of the FIM and FIM+FAM should have been administered at all sites. This was not possible for practical reasons. Therefore, results based on these analyses may have more limited generalisability.

The validation of the short-form FIM-8 was not undertaken in an independent sample from that used for item reduction and scale development. However, some of the psychometric properties of the FIM-8 have been examined subsequently in an independent sample of patients from the NRU (unpublished data). Similar results (data not reported) were obtained for internal consistency, correlations between scales, convergent validity (correlations with the Barthel Index and EDSS), discriminant validity (correlations with age and sex), and responsiveness. Nevertheless, the psychometric results of the short-form FIM-8 are preliminary and need to be confirmed in an independent sample.

6.3 Implications for clinical practice and research

6.3.1 Implications for clinical practice

Results from this study have several implications for clinical practice and service delivery. First, there is now good evidence that the FIM, FIM+FAM, and the Barthel Index are psychometrically sound measures of disability.

Second, results provide new information about which FIM and FIM+FAM scores should be reported. Third, results help to inform choice between the FIM, FIM+FAM, and Barthel Index. Finally, results call into question the validity of clinically derived conceptual models of disability.

This study confirms the scientific value of the FIM, FIM+FAM, and Barthel Index as measures of disability in neurorehabilitation, in addition to their previously agreed practical utility. All three measures have been incorporated successfully into the routine clinical practice of three busy units and have been shown to be scientifically sound. Consequently, they can be used for local audit, to compare outcomes between different units, and to evaluate the effects of policy changes.

Evidence from this study indicates that only total, motor, and cognitive scores should be reported for the FIM and FIM+FAM, as these are the only scores that have been shown to be reliable and valid. To date, there has been no consensus as to which FIM or FIM+FAM scores should be reported. Consequently, previous studies report various combinations of the whole range of scores (item, subscale, scales, total) for either groups or individual patients. Although subscale and item scores provide useful qualitative clinical information, there is insufficient evidence to support them as quantitative estimates of disability. Subscales were not shown to be valid measures of distinct dimensions of disability, even though they represent

reliable item groups. Reporting scores for single items is not recommended as they are unreliable and lack measurement precision and validity (100, 217, 269, 320).

Although the question of which FIM+FAM scores should be reported has not been addressed, it has been suggested that total scores should not be reported for the FIM (317). Using Rasch measurement techniques, Linacre *et al.* (321) showed that FIM items measure two distinct aspects of disability, motor and cognitive function. On this basis, they argued that the FIM is a multidimensional instrument, and that the FIM total score should not be reported as it is an ambiguous quantitative summary of disability which combines two distinct dimensions.

The fact that an instrument has multiple dimensions does not preclude a total score being reported, provided it can be justified on conceptual and empirical grounds (269, 271, 277). Indeed, most instruments designed to measure broad constructs like disability can be shown to have multiple subdimensions when subjected to statistical analyses of dimensionality (factor analysis, item-response theory, or log-linear models) (269, 271). This finding indicates, in a statistical sense, that the items of an instrument form multiple clusters based on the relative strengths of relations among them (269). However, it does not determine whether the instrument is measuring multiple constructs that are distinct or a single construct that is

heterogeneous (269, 277). Further examination of reliability and construct validity is required to make this distinction.

Although results from this study confirm Linacre *et al.'s* findings that the motor and cognitive domains of the FIM are distinct, they also indicate that reporting of FIM (and FIM+FAM) total scores is justifiable on conceptual and empirical grounds. Conceptually, it makes sense to combine all items to generate a total score as the different dimensions within the FIM and FIM+FAM are components of the construct of disability. Empirically, it is legitimate to report total scores as this study demonstrates them to be reliable and valid. It should be noted that Linacre *et al.* did not examine the reliability or validity of FIM scores.

Whether investigators should report FIM and FIM+FAM total or motor and cognitive scores depends on the purpose of measurement. Investigators wishing to examine motor and cognitive disability separately, or study the relationships between them, are more likely to report scale scores. Similarly, those interested in overall disability are more likely to report total scores. It is important to note that, for unequivocal interpretation of results, the construct measured must be homogeneous (271). Investigators should bear this in mind when analysing and interpreting FIM and FIM+FAM total scores.

Although health outcome measures are most commonly used to evaluate the performance of groups (51), an interesting question for further study is the appropriateness of using FIM and FIM+FAM scores for individual-patient clinical decision-making. This would be particularly useful in rehabilitation, which is a goal-oriented individually-tailored therapeutic intervention (322-324).

McHorney and Tarlov (266) discussed the issue of instruments used to assess individual-patients. They proposed six measurement standards: brevity, breadth and depth of health measured, cross-sectional and longitudinal measurement precision, and validity for individual-patient applications. Results from this study demonstrate that the FIM and FIM+FAM satisfy statistical standards for cross-sectional (alpha coefficients $\geq$ .90 - .95) and longitudinal (reproducibility $\geq$ .90 - .95) measurement precision. However, confidence intervals around individual-patient scores are extremely wide indicating that measurements are not accurate enough to be used to make clinical decisions about individual patients. Moreover, there are no data examining the validity of the FIM or FIM+FAM in an individual decision-making context. McHorney and Tarlov (266) examined the extent to which the SF-36, Functional Status Questionnaire (325), Dartmouth COOP Poster Charts (326), Nottingham Health profile (57), and the Duke Health Profile (327), met their six criteria as measures for individual decision-making. They also demonstrated wide confidence

intervals and the absence of relevant evidence for validity, and reached the same conclusions that these instruments were not appropriate in this context.

Results from this study allow clinicians to make an informed choice between the FIM, FIM+FAM, and Barthel Index. The fact that all three instruments are psychometrically comparable as evaluative measures indicates that no advantage is gained by selecting the longer instruments, the FIM and FIM+FAM, over the shorter Barthel Index. The choice of instruments, therefore, should be guided by other practical criteria. For example, the Barthel Index is cheaper and simpler than the FIM and FIM+FAM and does not require either trained raters or the use of team consensus to generate ratings. Consequently, the Barthel Index can be used easily in clinical settings where staffing is limited, for example outpatient clinics and domiciliary visits, enabling change in disability to be easily monitored over time following a period of hospitalisation. Furthermore, preliminary evidence (328, 329) for the validity of a postal version of the Barthel Index (232) suggests the possibility of ongoing disability measurement at minimum inconvenience to patients.

There are, on the other hand, four advantages of the FIM and FIM+FAM compared with the Barthel Index. First, they have superior measurement precision to the Barthel Index, indicating better ability to discriminate

between groups of patients. Second, the FIM and FIM+FAM provide specific information concerning motor and cognitive dimensions of disability, whereas the Barthel Index only addresses physical function. Third, as they have more items, the FIM and FIM+FAM provide a more extensive qualitative assessment of the patterns and areas of disability for individual patients and groups. That is, the FIM and FIM+FAM generate superior diagnostic information than the Barthel Index. Fourth, the process of team consensus rating enables dissemination of information about patients among the multidisciplinary team. Although no studies have examined this aspect of the rehabilitation process, therapists comment that these meetings are valuable for patient management.

Finally, results from this study indicate limited empirical support for clinically derived conceptual models of disability. Whilst the rationale underlying the development of the FIM and FIM+FAM is intuitively sound, the demonstration of their psychometric equivalence with the Barthel Index and failure to find empirical support for conceptual models indicates that disability is more complex than originally conceptualised. This area will be discussed later in this chapter.

This study also provides evidence for the effectiveness of inpatient neurorehabilitation. In the total sample and stroke patients, statistically significant improvements were shown for overall, motor, and cognitive

scores, whereas in MS patients, statistically significant improvements were demonstrated for overall and motor, but not cognitive scores. Although this is not a controlled study, these results provide evidence for the effectiveness of an intervention whose scientific basis is not very sound.

6.3.2 Implications for research

Results from this study have several implications for research. First, findings confirm the scientific rigour and thus usefulness of the FIM, FIM+FAM, and Barthel Index for research in neurorehabilitation. Second, results demonstrate the need for a more systematic and rigorous approach to the psychometric evaluation of existing instruments, and the development of new outcome measures before they can be recommended for use in clinical practice. Third, findings indicate the need for better conceptual models of disability. Finally, findings indicate the need for consensus about how responsiveness should be measured.

Findings from this study demonstrate that the FIM, FIM+FAM, and Barthel Index are all suitable measures for research. All three instruments satisfy standard criteria for reliability and validity and demonstrate responsiveness, indicating that they are appropriate for research applications such as randomised controlled clinical trials or observational studies. Moreover, the fact that they can be easily incorporated into clinical practice makes them

ideal for health services research. These findings increase the scientific credibility of results from studies, such as those examining the effectiveness of neurorehabilitation in MS (323, 330), that have used the FIM or FIM+FAM as disability outcomes measures in similar samples.

The psychometric equivalence of the FIM, FIM+FAM, and Barthel Index demonstrated in this study indicates that there is little advantage in using the two longer measures rather than the shorter Barthel Index. There are two caveats to this statement. First, it is necessary to use the FIM and FIM+FAM if separate motor and cognitive scores are required. Second, the superior measurement precision of the longer measures indicates an advantage over the Barthel Index if the purpose of the study is to examine group differences in disability. However, as most studies use health outcome measures for evaluative purposes, there is no psychometric advantage of one instrument compared with the others. Therefore, the choice of instrument must be based on other criteria. The practical advantages of the Barthel Index, discussed above, suggest it would be more successfully incorporated into large multi-centre studies. In addition, the availability of the self-report Barthel Index makes it possible to measure outcomes by postal survey. This is important for neurological disorders which are uncommon and disabling: postal survey allows the study of large samples of patients in geographically disparate locations and, by avoiding hospital attendance, reduces patient inconvenience, staff burden, and

cost. Alternatively, the FIM or FIM+FAM would be preferred for studies specifically examining motor and cognitive disability.

6.3.2.1 Evaluating existing health outcome measures

Results from this study indicate that the psychometric evaluation of existing measures should consist of a comprehensive examination of the full spectrum of measurement properties, the routine examination of incremental validity and, for multi-item measures, a detailed item analyses.

Many instruments used in clinical practice and research have been in existence for some years, and were developed before psychometric methods became familiar to clinicians. For example, the EDSS, the measure of disability in MS used to evaluate the effectiveness of interferons, was developed in 1954 (83). Similarly, the Rankin Scale, widely used to evaluate treatment effectiveness in stroke (331), was developed in 1957 (39). Likewise, the Barthel Index, probably the most widely used generic measure of disability in the UK, was developed in 1955 (65, 220).

In the last decade, the increasing awareness of psychometric methods, combined with recognition that the choice of outcome measure is crucial to the design of a successful clinical trial (70), has resulted in a flurry of

studies reporting *post hoc* psychometric evaluations. Whilst studies of individual measurement properties are increasingly common, there are few comprehensive evaluations of the full range of psychometric properties. Such studies are necessary because psychometric properties are largely independent of each other, and dependent on the samples in which they are examined. For example, the evaluation only of the reliability of an instrument (231, 238, 332) is of limited value in choosing an outcome measure for use in research. Although reliability is a pre-requisite for validity (91), it is possible to have a highly reliable scale with limited validity and poor responsiveness.

One property of instruments that receives little attention is acceptability. Although studies usually report mean or median scores, it is less common to report ranges of score, standard deviations or percentiles, and rare to report floor and ceiling effects or skewness statistics. As existing measures are often used in samples that differ widely, it is important that these data are reported to indicate the applicability of the instrument to the study sample. An example of the impact of unacceptable score distributions on the interpretation of results was discussed earlier in this chapter.

Acceptability data also provide useful information about the extent to which an instrument may achieve the aims for which it is being used. For example, the EDSS was developed specifically to detect between-group differences

and within-group changes in disability for MS patients (82). Acceptability data from this study show that, even though the EDSS mean score is situated near the scale mid-point and there are no floor or ceiling effects, there is little score variability. These results suggest that the EDSS has a poor ability to discriminate between individuals in terms of their disability, and suggest that its responsiveness will be limited. Further analyses have confirmed these predictions: each EDSS score is associated with a wide range of FIM and Barthel Index scores; EDSS measurement precision is 56% of the FIM total score; the effect size is .06.

Results from this study indicate that the routine evaluation of existing measures should go beyond a comprehensive assessment of their basic psychometric properties. The demonstration in this study that the FIM, FIM+FAM, and Barthel Index are all psychometrically sound but equivalent points to the need for including a comprehensive examination of incremental validity when evaluating all instruments. Whilst studies comparing individual psychometric properties and the use of instruments are increasingly common (23, 191, 262, 333-339), comprehensive head-to-head comparisons of instruments are not frequently undertaken. This may be because none of the current standards for evaluating instruments recommend an examination of incremental validity. However, an informed decision as to which of a group of measures is best for a study is dependent upon empirical evidence of their relative scientific and clinical properties.

Results from this study also indicate that the routine psychometric evaluation of existing multi-item measures should include item analyses. These are not frequently undertaken. For example, item analyses have not been reported for the Barthel Index, EADL, or OPCS. The value of item analyses is demonstrated by the fact that although the FIM and FIM+FAM satisfy standard criteria for reliability and validity, they show item redundancy, subscale overlap, and limited item discriminant validity. Here, item analyses indicate that both measures can be improved in their clinical usefulness (reduction of item number) and scientific rigour (more valid item groups). The short-form FIM-8 demonstrates that rigorous disability measurement can be achieved using a smaller number of items. Furthermore, item analyses demonstrated a misconceptualisation of disability which is discussed later in this chapter. Finally, and perhaps most importantly, item analysis of the FIM and FIM+FAM highlights some of the limitations of relying solely on basic tests of reliability and validity when evaluating instruments that were not developed using psychometric methods. For these reasons, item analyses should be part of the routine evaluation of available measures.

6.3.2.2 Developing new health outcome measures

Results from this study have implications for the development and evaluation of new health outcome measures before their introduction into clinical practice. They illustrate the limitations of clinical approaches to instrument development, and the importance of an iterative psychometric approach in which conceptual and measurement models are developed, item analysis and item reduction techniques are applied, the full spectrum of psychometric properties are comprehensively evaluated, and incremental validity is examined.

This study demonstrates that clinical approaches to measurement are not always supported by empirical results. The FIM was specifically developed because the Barthel Index was considered inadequate to evaluate the effectiveness of medical rehabilitation. Similarly, the FIM+FAM was developed because the FIM was considered inadequate to evaluate medical rehabilitation in patients with neurological disability. Although the method of development of both instruments was intuitively sound - items were selected by the consensus opinion of a panel of expert rehabilitation clinicians - empirical data from this study indicate that this approach has not achieved the desired goals as all three measures are psychometrically equivalent.

The finding of psychometric equivalence for the Barthel Index, FIM, and FIM+FAM highlights a difference between clinical assessment and measurement. There is no doubt that the greater number of items and response categories in the FIM and FIM+FAM provide more comprehensive qualitative assessments of disability than the Barthel Index. This probably explains why clinicians, who manage the day-to-day problems of individual patients, anecdotally prefer the FIM or FIM+FAM. However, all three instruments generate similar quantitative estimates of disability, largely because these additional items are redundant. For example, the four transfer items of the FIM+FAM are highly correlated indicating that, despite their clinical relevance to the functioning of individual patients, only one of these items needs to be included in a measure of disability. Although the finding that all transfer items are highly related may be predicted on clinical grounds, strong relationships between other items are less predictable. For example, transferring into a shower or bath is highly correlated with dressing, bathing, and toileting, indicating that this transferring item can be used in a measure to represent a wide range of clinical activities. This unpredictability concerning the extent to which items are related indicates that the selection of items for measurement is dependent upon the knowledge of their empirical relationships as well as their clinical relevance.

Similarly, the manner in which items are grouped into scales should not be determined by clinical intuition alone. Empirically derived item groups should also be generated. Then, the reliability and validity of both methods

of grouping items should be examined to determine the most valid operational definition of the construct being measured.

The nature of the construct of disability confirms the necessity for a psychometric approach to scale development. Disability, like many health constructs, is complex and based on theory. Disability measures operationalise this theory. For a construct to be measured rigorously, there needs to be empirical support for its conceptual basis. However, the development of measures for complex constructs often has to account for uncertainty about the underlying conceptual basis as well as the best method of assessment (243, 269). Psychometric methods recognise these uncertainties and the need to generate the most valid operational definition of the construct. Therefore, an iterative approach to scale development is advised in which conceptual and measurement models are proposed, evaluated, and refined on the basis of empirical data (243, 269). Subsequently, and in independent samples representative of those in which they will be used, instruments developed in this manner are comprehensively evaluated for their acceptability, reliability, validity, and responsiveness. Finally, as demonstrated above, incremental validity should be examined to provide empirical evidence of the advantages (or disadvantages) of one measure compared with alternatives. Had the developers of the FIM and FIM+FAM adopted a psychometric rather than a clinical approach to scale construction, the limitations of both instruments would have been identified and rectified during the development process.

For health care evaluation in neurology a new generation of instruments with superior measurement properties is required to meet the future clinical challenges. Advances in basic neurosciences have increased our understanding of the mechanisms of neurological disorders and led to the development and introduction of new therapeutic interventions. For example, approximately 30 disease modifying drugs are currently being evaluated for the treatment of MS (19). As their relative effects are likely to be marginal, accurate evaluation is dependent upon high quality measurement of health outcomes.

6.3.2.3 Refining conceptual models of disability

Results from this study have implications for the future of disability measurement as they raise concerns over the validity of current conceptual models of disability. Whilst these findings support the results of other studies, they do not clarify how disability should be measured and, therefore, further work is needed to refine conceptual models of disability.

The process of instrument development and evaluation can advance understanding of the conceptual basis of complex health constructs like disability. The items of a scale and the way they are grouped into

subscales, known as its measurement model, form an operational definition of the construct the instrument is intended to measure. The finding that empirical evidence does not support the measurement models of an instrument indicates a misconceptualisation of the construct being measured and provides a stimulus for further work on conceptual models of disability.

The measurement models of the FIM and FIM+FAM are based on clinical consensus. This has been the most frequently used strategy for selecting and grouping items of disability measures (340). As functioning involves purposeful activities, there has been a tendency to group items according to a common purpose or domain. Logically, this approach has resulted in domains such as basic or personal activities of daily living (ADL), instrumental or domestic activities of daily living (IADL), self-care, mobility, housework, social life, and recreation (341). Understandably, measurement models for disability measures, such as the FIM and FIM+FAM, are based on this intuitively sound conceptual model.

When this conceptual model is used as the basis for measurement, however, two assumptions must be satisfied. First, it must be demonstrated that different activities, represented as items, can indeed be combined as a measure and, also, provide unique information about disability. Second, different domains of disability, represented as subscales, must be shown to represent distinct aspects of disability as

domains that are highly related are confounded. Logically, activities (items) that are shown to be independent statistically cannot be considered indicators of the same domain of disability (subscale). Similarly, activities that are highly related statistically cannot be used to measure different domains of disability. In addition, activities that are very highly related provide almost identical information and, therefore, it is not necessary to include both items in the measure. The results generated by an instrument are most interpretable when its domains are both sensitive and specific: that is, each domain only measures the intended concept and discriminates this concept from those measured by the other domains of the instrument (127).

The measurement models of the FIM and FIM+FAM, although intuitively sound, are not supported by the findings of this study. Results demonstrate that activities hypothesised to be distinct (e.g. bathing, dressing, toileting, and the different types of transfers) are very highly related. Thus, they are not distinct and do not contribute to an instrument unique information about disability. Similarly, results demonstrate that domains of disability hypothesised to be distinct (e.g. self-care and transfer, communication and social cognition) are not. Furthermore, results demonstrate that activities thought to be representing distinct domains (e.g. community mobility representing locomotion and employability representing psychosocial adjustment), and therefore included in different subscales, are in fact highly related to multiple domains and, therefore, are neither specific nor

sensitive. The consequences of these misconceptualisations of disability are item redundancy, subscale overlap, and limited item-discriminant validity. Finally, the fact that a principal components analysis groups activities differently than that predicted by the developers (e.g. upper and lower limb function), provides further evidence of inadequacies in the conceptual models underlying the development of the FIM and FIM+FAM.

Other evidence which leads to a questioning of the measurement model of the FIM is provided by Stineman *et al.* (206). Their objective was "to seek more fine-grained impairment-specific dimensions beyond the motor and cognitive dimensions of the FIM" (page 636). They performed an item-level exploratory factor analysis (principal components method, varimax rotation) on FIM data for 20 different impairment categories including seven neurological impairments. Five different factor solutions were generated: one 2-factor solution (motor and cognitive disability); two 3-factor solutions (ADL, mobility, cognitive disability; upper cord, lower cord, cognitive disability), and two 4-factor solutions (self-care, sphincter, mobility, cognitive disability; low energy, high energy, sphincter, cognitive disability). Whilst the authors conclude that these findings indicate that the FIM is a multilayered, multidimensional measure of human function and that the impairment-specific subscales provide improved measurement, their results call into question the measurement model underlying the FIM. None of the factor solutions they obtained support the grouping of items into the

six domains hypothesised by the developers (self-care, sphincter, transfer, mobility, communication and social cognition).

These studies are not the first to demonstrate that empirical evidence fails to support clinically based conceptual models of disability. In fact, twenty years ago the Health Insurance Experiment (341) questioned the appropriateness of this method of item selection and grouping for measuring functional activities. Jette (340) was the first to address their question. Rather than examine the measurement models of specific instruments Jette performed a factor analysis (method not stated) of 34 widely used self-report ADL items. From a sample of approximately 200 elder adults with polyarticular disability (mean age 69), five factors were extracted that accounted for 58.5% of the total variance: physical mobility (10-items), kitchen chores (7-items), personal care (8-items), home chores (7-items), and transfer (2-items). Jette's results suggest that the concept of ADL is more complex than initially thought.

Recently, two studies (176, 342) have examined whether the hypothesised concepts of ADL and IADL are empirically distinct. Thomas *et al.* (342) performed an item-level exploratory factor analysis (principal axis method, oblique rotation) on the 14 items (7 ADL, 7 IADL) of a modified version of the Older Americans Research Survey (OARS). From a large data set (n = 8900; age > 65) three factors were extracted: basic self-care (5 items),

intermediate self-care (6 items), and complex self-management (3 items). Kempen *et al.* (176) factor analysed (principal component method, varimax rotation) the 18 items (10 ADL, 8 IADL) of the Groningen Activity Restriction Scale (GARS). Three factors were extracted with Eigenvalues exceeding 1.0: moderately difficult activities (9-items); simple activities (7-items); activities requiring special training (2-items). The findings from these studies indicate that ADL and IADL are not the distinct unidimensional constructs of disability that has been assumed.

While the studies of Jette, Thomas, Stineman, and Kempen indicate misconceptualisations about disability, they do not clarify its conceptual basis or how it should be measured. There are two reasons for this. First, none of these studies has examined the validity of the item groups generated. Although factor analysis is widely used to define how the items of an instrument might be grouped into scales, examination of the construct validity of these item groups is essential to determine how they should be interpreted. It was noted earlier that when item pools representing diverse constructs such as disability are subjected to examinations of their dimensionality, multiple item clusters are often demonstrated. Whether these clusters of items represent distinct dimensions of measurement, or simply reflect the heterogeneity of the construct, remains unresolved until intercorrelations between the scales and item and scale convergent and discriminant validity are examined.

The second reason that these studies do not clarify the conceptual basis of disability is that none of them has defined the construct being measured. All four studies discussed above generate different item groupings. This is not surprising as the solutions generated by factor analysis are dependent upon the initial item pools analysed (305), which were unique for each study. To determine the conceptual basis of a construct, it must first be defined clearly to ensure that all important items are considered for inclusion (271). If a construct is not defined carefully in advance, the extent to which the items analysed are representative of the domain of interest remains uncertain, and connection between the construct and the scale is unclear (269).

Work is still required to determine empirically based conceptual models of disability. This should be undertaken as a collaboration between clinicians and measurement experts. Future disability scales need to be based on clear definitions of the aspect/s of disability that investigators are attempting to measure. Large item pools should be generated from semi-structured patient interviews, review of the literature and existing measures, and expert clinical opinion. Measurement models should be developed, tested, and refined accordingly.

6.3.2.4 Evaluating responsiveness

Results from this study highlight the need for consensus about how responsiveness should be reported, question the use of Cohen's criteria in the clinical interpretation of responsiveness data, and demonstrate the importance of examining relative responsiveness. Until a consensus is achieved, it is suggested that an hypothesis testing approach to examining responsiveness, akin to gathering evidence for validity, might be appropriate.

There is an urgent need for consensus about how responsiveness should be reported. As demonstrated in this study, five commonly used methods of calculating responsiveness generate estimates of different magnitude. Although this finding is predictable (each method has a different statistical formula), if investigators are not aware of this fact it can lead to different interpretations of instrument responsiveness. If, however, there are consistent relationships between the magnitude of the estimates generated by different statistics interpretation is more straightforward. For example, if estimates generated by standardised response means are consistently twice the magnitude of estimates generated by effect sizes (as they are in this study) one statistic can be predicted the another. Unfortunately, empirical evidence demonstrates that this is not the case. For example, results from other responsiveness studies demonstrate that the magnitude of estimates generated by standardised response means can be greater than (343),

equal to (236, 291, 343, 344), or less than (318, 344) the magnitude of responsiveness estimates generated using effect sizes.

Another problem that arises from the different responsiveness estimates generated by effect sizes and standardised response means concerns their clinical interpretation. Cohen's criteria (.2 = small; .5 = medium; .8 = large (311)), are generally used for the clinical interpretation of responsiveness estimates generated by these two methods (132, 135). However, findings from this study demonstrate that quite different conclusions can be drawn from the same data: the responsiveness of the FIM total score is medium in terms of the effect size and large in terms of the standardised response mean.

A further reason for not recommending the use of Cohen's criteria is that responsiveness estimates are dependent on the intervention and the disease studied. In this study, responsiveness coefficients for the FIM and FIM+FAM are higher in stroke than MS patients suggesting that these instruments are more responsive in stroke patients. However, as responsiveness coefficients are standardised change scores, and as FIM and FIM+FAM change scores are expected on clinical grounds to be smaller in MS than stroke patients, it is predictable that MS patients will have smaller responsiveness estimates. These findings by no means indicate that the FIM and FIM+FAM are less able to detect change in disability for

MS patients, and support the findings of other studies demonstrating that responsiveness is dependent on the magnitude of the change induced by an intervention (138, 291, 345). Moreover, this study demonstrates that the magnitude of the change induced by rehabilitation is disease-dependent. The responsiveness of an instrument, therefore, reflects three variables: the ability of the instrument to detect change, the ability of the intervention to induce change, and the potential of patients in the sample to undergo change. Consequently, responsiveness cannot be viewed as a property of the instrument itself.

Although sample dependency is a feature of all psychometric properties, it is notable in this study that responsiveness appears more sample dependent than reliability and validity. Furthermore, results from this study provide evidence that responsiveness, unlike reliability (13), is not related to the number of items in a scale (272) or the number of item-response categories (346, 347). These findings emphasise the importance of examining the relative responsiveness of different instruments in the same clinical setting. Using this method, the study sample and therapeutic intervention are held constant, and the results generated reflect the relative ability of the different instruments to detect change in that specific clinical setting.

The sample dependency of responsiveness coefficients calculated from pre- and post-intervention change scores, defined as prospective methods of determining responsiveness (138), has prompted investigators to seek other methods of estimating the ability of instruments to detect change. One commonly used approach is to compare change scores and an external criterion of change, such as a transition question (260). In this method, patients or clinicians assess the amount of change retrospectively using a global scale of change (e.g. -3 = markedly worse, 0 = no change, +3 = marked improvement). Responsiveness can then be determined in a number of ways. For example, by correlating change scores with scores from the global measure of change (high correlations indicate greater responsiveness) (131). Alternatively, the minimum clinically important difference (mean change score for minimally improved / deteriorated patients minus mean change score for unchanged patients) can be calculated (348). Or, the coefficient proposed by Guyatt *et al.* (110) can be calculated (mean change score in patients judged to have changed divided by the standard deviation of change scores in patients judged to have not changed). These methods have been defined as retrospective methods of determining responsiveness (138).

Recently, Norman *et al.* (138) have compared prospective and retrospective methods of examining instrument responsiveness. Using data from several studies and simulation methods, the authors investigated the relationship between responsiveness estimated using standardised response means

(prospective) and estimates calculated using Guyatt's method (retrospective). They demonstrate that there is no predictable relationship between the two indices either within or across studies. The authors indicate that the root of the problems is the confounding of natural variation among patients with experimentally induced change: an effect first anticipated by Cronbach forty years previously when he described two contrasting paradigms within psychology (349). He differentiated the correlational approach, concerned with the study of individual differences in mental capacity, from the experimental approach, concerned with controlling and predicting behaviour through experimental interventions.

Norman *et al.* reason that the two approaches to estimating responsiveness correspond to these paradigms. Responsiveness estimated prospectively is consistent with the experimental paradigm as the magnitude of the statistic relates the size of the treatment effect (numerator) to the variation in individual treatment response. Responsiveness estimated retrospectively is consistent with the correlational paradigm as it is based on variations between subjects in response to treatment, and has no direct relationship to the overall treatment effect. They conclude that retrospective methods of computing responsiveness yield little information about the ability of an instrument to detect treatment effects and recommend that they are not used as a basis for selecting instruments for clinical trials.

Determining instrument responsiveness is, therefore, complex. On the one hand prospective methods cannot divorce the ability of the instrument to detect change from the magnitude of the change induced by the intervention. On the other hand retrospective methods appear to yield little information about the ability of an instrument to detect treatment effects. Until consensus is reached as to how responsiveness should be measured, an hypothesis testing approach similar to that used to gather evidence for the validity of an instrument might be an appropriate strategy. In this method, the responsiveness of an instrument is determined in multiple studies designed to examine the extent to which the instrument is able to detect change in the construct being measured.

The sample and intervention dependency of prospective methods of examining responsiveness can be used advantageously. Evidence supporting the responsiveness of an instrument is provided if hypotheses concerning differential responsiveness are confirmed empirically. For example, patients with relapsing-remitting MS are hypothesised to make greater functional gains from rehabilitation than patients with the progressive forms of this disease because of more extensive spontaneous neurological recovery. Recently, this hypothesis has been confirmed for the FIM and Barthel Index (213). Similarly, the fact that patients with secondary progressive MS undergo a faster deterioration in disability over time than patients with primary progressive MS could be exploited to provide evidence of responsiveness. Furthermore, the finding in this study that effect sizes

are greater for the motor scales than the cognitive scales of the FIM and FIM+FAM are consistent with clinical predictions as rehabilitation predominantly effects physical function. Consequently, the demonstration that instruments are able to detect different degrees of change provides support for their ability to detect change.

Despite Norman *et al.'s* (138) advice not to use retrospective methods to quantify instrument responsiveness, these methods can be used to provide evidence of responsiveness. For example in this study, the demonstration that increases in staff ratings of improvement in disability on discharge are associated with a stepwise increase in magnitude of change scores cannot be ignored as support for the responsiveness of the FIM and FIM+FAM. Other authors (318) have demonstrated similar findings. It is notable that using retrospective methods in this manner differs from Norman *et al.* who studied the relationship between responsiveness quantified using prospective and retrospective methods.

Using the approach outlined above the extent to which empirical evidence of responsiveness gathered from multiple studies supports *a priori* predictions based on clinical expectation can be used to provide strong evidence for the responsiveness of measures despite the limitations of the inherent methods. Clearly, more work is required to investigate the potential of this approach to responsiveness testing.

6.4 Future directions

6.4.1 Clinical acceptability of the FIM, FIM+FAM and Barthel Index

Work is required to determine the extent to which the FIM, FIM+FAM, and Barthel Index are rigorous measures of disability for neurorehabilitation in the UK. Although results from this study provide strong support for their clinical usefulness and scientific soundness, and even though all three measures are widely used across the UK, there are few data examining their psychometric properties in other samples of rehabilitation patients. In fact, other studies using the FIM, FIM+FAM, and Barthel Index in rehabilitation settings report findings that question their acceptability (191, 214, 216). A study examining the score distributions of the FIM, FIM+FAM, and Barthel Index in rehabilitation units that routinely collect these data would provide this useful information. In addition, within-scale analyses could be undertaken on these data (e.g. alpha coefficients, item-total correlations, inter-item correlations, and intercorrelations between scales) to examine further their psychometric properties in different samples.

If data confirm the widespread scientific soundness of the FIM, FIM+FAM, and Barthel Index in UK rehabilitation centres, work is then needed to

generate normative data to enable the interpretation of scores, in particular the meaning of change scores. Although the UDS collate FIM scores from subscribers, these data are not freely available. Otherwise, few normative data exist (350) and there are none for the UK. These data are required to guide patient selection, to evaluate different rehabilitation practices, to compare the effectiveness of different interventions, and to guide sample size calculations for clinical trials. Defining poor outcomes would help identify patient groups that might benefit less from rehabilitation. For these groups alternative management strategies are more appropriate, such as maximising the home environment and improving community services. At present, these decisions are based on the experience of rehabilitation clinicians. As no two rehabilitation units have the same clinical practice, evaluation of outcomes will enable comparisons of the relative contributions of the individual components of the rehabilitation process to be defined. In addition, there have been few comparisons of rehabilitation with other therapeutic interventions. Finally, sample size calculations for clinical trials require a meaningful description of effect sizes (293).

6.4.2 Interpreting scores generated by health outcome measures

Despite the widespread use of health outcome measures, there is no systematic strategy for translating health outcomes into clinical decisions. Within psychology, it is commonplace to relate scale values and change

scores to the rest of the sample (or ideally, to the population) in terms of standard deviation units, percentiles, or percentages. These methods enable comparisons across samples, constructs, and measures. Although these statistical benchmarks are importance for interpreting scores, they are unfamiliar to clinicians and their patients and, more importantly, have limited clinical meaning. Moreover, statistical significance does not ensure clinical significance (and *vice versa*) (293).

Content based referencing is required to give clinical meaning to change scores (127, 351). That is, changes on health measures need to be anchored to clinical or other relevant changes (352). Whilst clinical interpretation of scores is relatively straightforward for single item measures like the EDSS - each score has a specific clinical meaning - it is less clear for multi-item measures as all scores (except the minimum and maximum) represent multiple permutations of item scores. Even determining the clinical significance of changes on single item measures is not simple. For example, the recently published randomised placebo controlled study of interferon beta in secondary progressive MS (59) demonstrated a statistically significant finding of less disability in the treatment group. The difference of .13 EDSS points, less than half of a level, may have significant health policy implications but the clinical significance has not been determined.

Several methods have been proposed to aid the clinical interpretation of Barthel Index and FIM scores. For the Barthel Index, scores in the 0 to 8 range indicate complete dependence, 8 to 12 partial dependence, and 12 to 20 full independence (228). For the FIM, Granger *et al.* have studied the relationship between FIM scores and the number of minutes help per day required from another person for patient with MS (208) and stroke (211). Results from these small studies (n = 20) demonstrate that a change of one FIM point is associated with a change in help required of 3.38 minutes in MS patients, and 2.19 minutes in stroke patients. Another method for the clinical interpretation of FIM scores has been to examine the score differences associated with different discharge destinations (e.g. independent at home, care at home, sheltered accommodation, residential care) (195).

Whilst these studies provide support for the external construct validity of the FIM (as predicted low scores are associated with more help in minutes and discharge to residential care), they are limited. For example, the amount of help disabled people receive is dictated by the amount of help available (e.g. family and social services) and influenced by the disablement friendliness of the environment. Indeed, these features influence self-reports of disability (353). Similarly, the discharge destination of patients may depend more on the skill and facilities of the clinicians involved in the placement process than the disability of the patients.

Several other methods have been proposed for the clinical interpretation of scores on health measures. These include relating scores or score changes to the cost of health care utilisation (354), major life events (355), preference weightings (356), equivalence with the impact of other diseases (357), or visual representations (mapping) of the relationships between perceptions and behaviours (358). All these methods have their limitations (352) (for example, major life events are uncommon and their impact is variable), prompting Deyo *et al.* (359) to recommend the use of a limited number of measures, and Lydick and Yawn (360) to add that the continued collection of data concerning clinical anchors will enable clinicians, over time, to become increasingly familiar with the clinical significance of particular levels of change.

One method for giving clinical meaning to change scores is Jaeschke *et al.'s* (348) minimum clinically important difference (MCID) approach. In this method, change scores are related to direct ratings of change generated by patients and/or clinicians on a transition question: the MCID is calculated as the mean change score for those persons rated as minimally improved minus the mean change score for those persons rated unchanged. Results from two studies (137, 348) suggest that, when using 7-point rating scales, the MCID for a measure is .5 scale units per item, a change of 1.0 units per item is considered moderate, and of 1.5 units per item is considered large. Although Juniper *et al.* (137) suggest that these changes are generalisable

to all areas of health-related quality of life, results from this study suggest otherwise. Table 6.3 reports, for stroke and MS patients, mean change scores for the FIM total scale for each level of staff-rated change in disability on discharge. Whilst the sample sizes are small and the calculation of an MCID is probably inappropriate, there is a two or three fold disease-related difference in the magnitude of the mean change score for each level of improvement.

These results are, perhaps, to be expected. The clinical significance of changes in health status are dependent on raters' expectations and patients' needs. For example, stroke patients are expected to achieve much greater functional gains from rehabilitation than MS patients. Similarly, different degrees of change in disability will have variable clinical significance for individuals. Moreover, transition questions (theoretically) have limited reliability, validity, and precision (138) as they are single item measures (271). If the MCID is to be used for the clinical interpretation of changes in health measures, the psychometric properties of transition questions must be comprehensively documented and disease-specific reference values are required.

6.4.3 Evaluating the effectiveness of neurorehabilitation

Although this study has evaluated the usefulness of two disability measures, the role of disability itself as an outcome of rehabilitation needs to be considered critically. To evaluate rigorously therapeutic effectiveness, investigators must decide and specify in advance what the intervention is trying to achieve. Based on this information, measures should be chosen or developed that comprehensively evaluate these domains of health. Whilst the WHO's International Classification of Impairments, Disabilities, and Handicaps (ICIDH) (155) has been the framework behind many studies of rehabilitation, further work is required to define a conceptual basis for outcome studies in rehabilitation. As shown in this study, psychometric methods can help to do this.

Rehabilitation is concerned with restoring, to whatever degree may be possible, an individual's capacity for integrated functioning in physical, psychological, and social terms (157). It is a multidisciplinary, problem-solving, goal-oriented, individually tailored intervention, the aim of which is to lessen the impact of the disease on daily life and to enable patients to realise their own potential within the limitations posed by their disease (361). Disability is just one of many aspects of health effected by this complex intervention and, therefore, comprehensive studies of rehabilitation effectiveness should also include an evaluation of other relevant health constructs.

The ICIDH (155) provides the framework for many studies of rehabilitation. This classification was developed as a response to the need for a conceptual and operational framework for describing and measuring the consequences, or disablement, of chronic conditions (157). The conceptual distinctions were developed specifically to correspond to the obligations of different components of health care: impairments were considered the concern of medical services, disabilities the concern of rehabilitation services, and handicaps the concern of social welfare provisions (157). It was thought that this classification would bring coherent thinking to an arbitrary and disjointed area, lead to better understanding and communication and, as a result, improved health care (362). The ICIDH has achieved its goal: it is considered to be the cornerstone of rehabilitation management as it is comprehensive, relevant and amenable to operationalisation, and each concept can be defined influenced and measured (156, 363). As a result, studies of rehabilitation have determined effectiveness by measuring one or more of impairment, disability, and handicap outcomes.

Although the ICIDH provides a conceptual model of disablement for evaluating the rehabilitation process, the empirical basis of the model has not been tested. Whilst it is widely accepted and clearly documented that the rehabilitation process addresses a vast number of diverse health-related

problems (150), there is uncertainty as to which dimensions of health are influenced. The situation is complicated further as clinicians and commissioners of health care do not agree on the aims of rehabilitation services. Commissioners often are looking for simple independence in ADL, discharge from hospital, or return to paid employment, whereas clinicians often have wider interests such as maximising leisure participation and social interaction (364). Moreover, clinicians do not agree about which outcomes should be measured. Wade reports from a recent international meeting, held in Vienna, to discuss outcomes measurement in rehabilitation: "some people wanted more disease-specific measures but others wanted more generic measures: some wanted longer more detailed measures while others wanted shorter, simpler measures; the level of measurement (impairment, or disability, or handicap) was debated" (364, page 93).

Wade's opinion is that too much emphasis is placed on outcome measurement instruments and not enough on study design. Using the analogy of workers (study design) and their tools (outcome measurement instruments), he argues that good workers (randomised controlled trials) using poor tools achieve better results (outcomes) than poor workers (observational studies) using good tools. This opinion is in direct contrast to the more accepted view expressed in the psychological measurement literature that training in measurement principles is an area of relative neglect compared to the emphasis on study design (81, page 653).

Moreover, whilst rigorous study designs are essential, Wade's view fails to account for three issues. First, there is a clear consensus that observational studies do indeed provide scientifically credible information and randomised controlled studies are subject to their own problems, such as limited generalisability (365-367). For example, it is notable that most studies of interferons in MS have excluded the majority of people with the disease. Furthermore, in some cases undertaking randomised controlled trials of inpatient rehabilitation may not be feasible as such studies would require control groups to be admitted to a rehabilitation unit for a sham intervention (323).

The second issue that Wade's opinion fails to account for is that any study has a higher scientific impact when the outcome measures used are proven to be rigorous. Finally, Wade's opinion fails to recognise that psychometric methods can be used to answer the question of which outcomes should be measured. The process of instrument development using psychometric methods, as discussed earlier in this chapter, can lead to advances in our understanding of the conceptual basis of health. Studies using psychometric methods to develop measures for evaluating rehabilitation outcomes, would result in the development of empirically based conceptual models of the impact of rehabilitation. These studies are urgently needed as the scientific basis of rehabilitation is not firmly grounded (149, 366).

It is important that new measures of neurorehabilitation, and neurology in general, consider the self-report method of administration. There are conceptual and methodological reasons for this. First, from a conceptual point of view, the person receiving an intervention is best placed to judge its benefit (49). Also, the perceptions of patients and the health professionals who treat them differ (41), and many important domains of health (e.g. emotional well-being) are subjective concepts (36). From a methodological point of view the self-report method of administration of health measures reduces the burden of outcomes data collection and enables postal surveys to be conducted. In neurology, as many disorders are uncommon, the recruitment of an adequately sized sample often requires multiple centres. Furthermore, many patients with neurological disorders are significantly disabled and hospital visits are often inconvenient, troublesome, and problematic. Finally, self-report methods of administration allow measurement instruments to be used in and out of hospital, thereby enabling the collection of better follow-up data.

Before the development of a wave of self-report measures for neurology and neurorehabilitation, work is needed to clarify the impact of motor and cognitive disabilities on this method of administration. Many neurological disorders are associated with cognitive impairment. For example, approximately 70% of people with MS have abnormal cognitive functioning on formal neuropsychological testing (253). The impact of cognitive impairment on questionnaire completion and psychometric properties is

poorly studied. Research is needed to quantify the levels of cognitive impairment associated with adequate questionnaire completion and the impact of individual deficits (e.g. memory, reasoning, and attention). Similarly, little is known about the impact of visual and upper limb dysfunction, either separately or together, on the completion of self-report questionnaires.

An area that requires further evaluation is the use of individual patient measures to evaluate the outcome of neurorehabilitation. Some authors (324, 368) argue that standardised measurement instruments, like the FIM, FIM+FAM, and Barthel Index are less relevant for measuring the outcomes of rehabilitation than other areas of medicine. This is because rehabilitation treats a heterogeneous mix of patients with a broad range of diseases and disabilities, and has diverse individually tailored treatment goals (324). Furthermore, others (209) argue that functional status measures are unable to record the small but clinically significant changes that often accompany successful rehabilitation.

Goal Attainment Scaling (368) is a widely used technique (322) which was developed to measure aspects of health specific to both the patient and the aims of the intervention. Its major advantage over other measurement methods in rehabilitation is that it compares a patient's actual achievement to what the patient could be expected to accomplish. Therefore, it is a direct

measure of patient change and treatment effectiveness. The major limitations of Goal Attainment Scaling are that highly specific training (at least one year of experience) is required to selected and scale the goals. Recently, Improvement Scaling (324) has been developed. This is based on Goal Attainment Scaling, but is more user friendly (briefer) and requires less training as the goals (usually three) are selected from a list of 65 standardised (pre-scaled) goals.

Preliminary data gathered using Improvement Scaling is encouraging (324). However, the feasibility of using individual patient assessment measures in routine clinical practice has been questioned because they are time consuming (369). More importantly, empirical evidence is required to determine whether this is a reliable and valid method of measurement. Finally, there are few guidelines as to how validity can be empirically determined; this is assumed on the basis of content validity and patient specificity. However, the role of Improvement Scaling to measure outcomes in rehabilitation requires further examination.

6.4.4 Item reduction techniques

In the development of multi-item measures, one area that requires further investigation is the comparison of item reduction techniques. Whilst there is consensus that items should be generated from multiple sources to ensure

that all important variables are considered for inclusion in a scale, and consensus that item reduction and scale formation should be empirically based, there is no consensus as to which item reduction method should be used. Recently, two studies (297, 370) have demonstrated that the two principal methods of item reduction and scale formation, psychometric and clinimetric, generate two different instruments from the same item pool. This finding has wide reaching implications; if measures of the same phenomena, developed using different methods, generate different results, the validity of one or both methods (and all studies using scales developed by that method) is questioned (370). The implications of these findings have yet to be fully determined.

The two methods of item reduction have different philosophies (297). In the clinimetric strategy (371), items are chosen from the item pool on the basis of empirical evidence of their importance to patients and/or clinicians. In the psychometric strategy, items are chosen on the basis of their empirical performance as measures.

In a retrospective study, Juniper *et al.* (297) compared clinical impact (clinimetric) and factor analysis (psychometric) methods of reducing 152 items in the development of an instrument to measure quality of life in adults with asthma. In the clinimetric method, patients indicated whether they had experienced each item during the last year (yes / no), and rated the

importance of each item on a five-point scale (1 = not important, 5 = extremely important). For each item, an impact score was calculated as the product of the proportion of patients experiencing the item and its mean importance score. Items with the highest impact scores were selected, and then grouped into scales on the basis of clinical opinion. The final questionnaire contained 32 items in four scales: symptoms, emotional function, activity limitation, and environmental exposure.

In the psychometric method, items experienced by less than 40% of patients were discarded. Redundant items (item intercorrelations > .70) were identified, and the item with the lowest item-total correlation was eliminated. A principal components analysis was conducted and items were removed that loaded on the first factor by less than .40. Last, using varimax rotations, solutions were generated with three, four, five, and six principal components, and the most clinically sensible of these solutions was retained. The final questionnaire contained 36 items in five scales: unpleasant chest sensations, fatigue and emotional function, activity limitations, symptoms of nocturnal asthma, and impairments associated with environmental stimuli. Therefore, the clinimetric and psychometric methods of item reduction generated instruments with a different number of items and scales and different scale content. Twenty items were present in both instruments. Juniper *et al.* conclude that the two methods of item reduction lead to appreciably different instruments.

Marx *et al.* (370) compared, prospectively, clinimetric and psychometric methods for selecting the best 30 items from a 70-item pool to generate a measure of upper extremity disability. In the clinimetric strategy, items were selected with the highest mean scores for aggregated importance and severity ratings, each measured on five-point scales (1 = not at all, 5 = extremely). In the psychometric strategy, items were selected with the highest equidiscriminatory item-total correlations (a modification of item-total correlations that identifies items that correlate with each other but discriminate between individuals throughout a range of scores (217)). Finally, clinicians modified both 30-item scales to improve their face validity by exchanging items from the rejected item pool. Ten items were exchanged in the clinimetric scale, three items in the psychometric scale. Before clinician modification fifteen items were common to both 30-item instruments, sixteen items post-modification.

Marx *et al.* examined further the similarities of the two instruments by computing alpha coefficients, mean total scores, and agreement between scores using the method of Bland and Altman (372) and an intraclass correlation coefficient. Results for clinimetric and psychometric scales, respectively, were: (before clinician modification) alphas = .96 and .97; mean score = 48.2 and 39.1; limits of agreement (mean difference $\pm$ SD) = 9.1 $\pm$ 17.6; ICC = .93; and (after clinician modification) mean score = 40.9 and 39.2; limits of agreement (mean difference $\pm$ SD) = 1.7 $\pm$ 10.4; ICC =

.97. Marx *et al.* conclude that although the scales differ in content, clinimetric and psychometric strategies for item reduction are complementary.

Whilst there is no doubt that the two methods generate instruments with different item content, the two studies reach somewhat different conclusions. The implications of psychometric and clinimetric methods of item reduction, therefore, remain unclear. Furthermore, there are limitations in both studies that influence the interpretation of results. Juniper *et al.*, although concluding that the instruments are different, do not examine the extent to which the two instruments are different from a psychometric point of view. Marx *et al.*, although concluding that the two methods of item reduction are complementary, do not report a comprehensive comparison of all psychometric properties of the instruments to determine fully the extent of their similarity. For example, a high correlation between the two measures is expected as item overlap is high (53%) and the range of scores is wide (0 to 100). It is also notable that whilst the clinimetric methods used in the two studies are very similar (items are selected with highest impact factor), they differ in terms of the psychometric methods used. Juniper *et al.* use an item elimination strategy based on four criteria. In contrast, Marx *et al.* select items on the basis of an entirely different criterion. The extent to which these differences in method influence the results is uncertain.

In practice, the distinction between psychometric and clinimetric methods may not be as black and white as portrayed in these two studies. Whilst psychometric methods are indeed more statistically based than clinimetric methods, it is strongly recommended that the approach is inductive rather than deductive (269). That is, scale development should begin with a clearly defined construct which guides subsequent scale development. Similarly, validation should be confirmatory with theoretical ideas guiding the validation strategy and hypotheses generated about the relationships between the scale and other variables. In contrast, deductive methods (e.g. factor analysis) allow the results to dictate the construct underlying the item pool. Factor analysis is just one of the many analyses used in item reduction and has many limitations (305, 309, and 271 page 533-535). Ultimately, scale development involves a combination of both clinimetric and psychometric methods to generate the most valid operational definition of a construct.

6.5 Concluding remarks

In a recent article, McDowell and Jenkinson (127) called for the academic discipline of health measurement to be developed further. There is no doubt that this is essential. Expressions of concern about health measurement instruments and the ways they are applied in clinical settings are increasing (54). Anecdotally, clinicians comment particularly on the frustration and

confusion they feel when asked to administer questionnaires termed "quality of life" measures that differ widely in content.

A formal discipline is required to raise awareness of clinicians and guide instrument choice. A UK panel of experts is required to develop consensus guidelines for the evaluation and development of new measures, to advise granting bodies and journal editors, and to co-ordinate research programmes. Further development of theory and conceptual models is required at the expense of questionnaire development (54).

The fundamental view of this thesis is that in health outcomes measurement scientific rigour should not be compromised. The impetus for this view is not only the advancement of science, the encouragement of genuine intellectual enquiry, or cost containment, but primarily the pursuit of better patient care.

References

1. Harris AI. Handicapped and impaired in Great Britain. Part 1. London: Office of Population Censuses and Surveys, 1971.

2. Wade DT. Epidemiology of disabling neurological disease: how and why does disability occur? Journal of Neurology, Neurosurgery, and Psychiatry 1996;61:242-249.

3. Wade DT, Langton Hewer R. Epidemiology of some neurological diseases. International Rehabilitation Medicine 1987;8:129-137.

4. Langhorne P, Dennis M, editors. Stroke units: an evidence based approach. London: British Medical Journal Books, 1998.

5. Hatch J. The economic impact of multiple sclerosis. MS Management 1996;3:40.

6. Jonsson B. The economic cost of multiple sclerosis in Sweden. Stockholm: Stockholm School of Economics (report number EFI-6551), 1995.

7. Prouse P, Ross-Smith K, Brill M, Singh M, Brennan P, Frank A. Community support for young physically handicapped people. Health Trends 1991;23:105-109.

8. Harvey C. Economic costs of multiple sclerosis: how much and who pays? London: National Multiple Sclerosis Society (report number ER-6005), 1995.

9. Dennis M. Stroke physicians unite to improve care. Hospital Doctor 1999:15th April; 34-35.

10. Campbell NR. Symposium: measurement and its importance for philosophy. Aristotelian Society Supplement 1938;17:277-334.

11. Stevens SS. On the theory of scales of measurement. Science 1946;103:677-680.

12. Torgerson WS. Theory and methods of scaling. New York: John Wiley and Sons, 1958.

13. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford University Press, 1995.

14. Bohrnstedt GW. Measurement. In: Rossi PH, Wright JD, Anderson AB, editors. Handbook of survey research. New York: Academic Press, 1983:69-121.

15. Bradford Hill A. Principles of medical statistics. 9th ed. London: The Lancet Limited, 1971.

16. Bunker JP, Wenneburg JE. Operation rates, mortality statistics, and the quality of life. New England Journal of Medicine 1973;289:1249-1251.

17. World Health Organisation. The constitution of the World Health Organisation. WHO Chronicles 1947;1:29.

18. van der Bos GAM, Limburg LCM. Public health and chronic diseases. European Journal of Public Health 1995;5:1-2.

19. Thompson AJ, Noseworthy JH. New treatment for multiple sclerosis: a clinical perspective. Current Opinion in Neurology 1996;9:187-198.

20. Frater A, Costain D. Any better? Outcome measures in medical audit. British Medical Journal 1992;304:519-520.

21. Reiser SJ. The era of the patient. Journal of the American Medical Association 1993;269:1012-1017.

22. Hobart JC, Freeman JA, Lamping DL. Physician and patient-oriented outcomes in chronic and progressive neurological disease: which to measure? Current Opinion in Neurology 1996;9:441-444.

23. Jenkinson C, Peto V, Fitzpatrick R, Greenhall R, Hyman N. 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 1995;24:505-509.

24. Fillipi M, Paty DW, Kappos L, Barkof F, Compston DA, Thompson AJ, Zhao GJ, Wiles CM, McDonald WI, Miller DH. Correlations between changes in disability and T2-weighted brain MRI activity in multiple sclerosis: a follow up study. Neurology 1995;45:255-260.

25. Smith D, Baker GA, Jacoby A, Chadwick DW. The contribution of the measurement of seizure severity to quality of life research. Quality of Life Research 1995;4:143-158.

26. Peto V, Jenkinson C, Fitzpatrick R, Greenhall R. The development and validation of a short measure of functioning and well-being for individuals with Parkinson's Disease. Quality of Life Research 1995;4:241-248.

27. The IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I Clinical results of a multi-centre,

randomised, double-blind, placebo-controlled trial. Neurology 1993;43:655-661.

28. Jacobs LD, Cookfair DL, Rudick RA, Herndon RM. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. Annals of Neurology 1996;39:285-294.

29. Ebers GC, Oger J, Paty D. The multiple sclerosis PRISMS study: prevention of relapses and disability by interferon beta-1a subcutaneously in multiple sclerosis. Annals of Neurology 1997;42:986.

30. McDonald WI. New treatments for multiple sclerosis. British Medical Journal 1995;310:345-346.

31. Mumford CJ. Beta interferon and multiple sclerosis: why the fuss? Quarterly Journal of Medicine 1996;89:1-3.

32. Anonymous. Interferon beta-1b in MS - hope or hype? Drug and Therapeutics Bulletin 1996;34:9-11.

33. Harvey P. Why interferon beta-1b was licensed is a mystery. British Medical Journal 1996;313:297-298 (letter).

34. Gill TM. Quality of life assessment. Journal of the Royal Society of Medicine 1995;88:680-682.

35. Ware JE Jr. The status of health assessment 1994. Annual Review of Public Health 1995;16:327-354.

36. Stewart AL, Ware JE Jr, editors. Measuring functioning and well-being: the Medical Outcomes Study approach. Durham, North Carolina: Duke University Press, 1992.

37. O'Donoghue MF, Duncan JS, Sander JWAS. The subjective handicap of epilepsy: a new approach to measuring treatment outcome. Brain 1998;121:317-343.

38. Kurtzke JF. Rating neurological impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983;33:1444-1452.

39. Rankin J. Cerebral vascular accidents in patients over the age of 60 : II. Prognosis. Scottish Medical Journal 1957;2:200-215.

40. Lang AET, Fahn S. Assessment of Parkinson's disease. In: Munsat TL, editor. Quantification of neurological deficit. Stoneham, Massachusetts: Butterworths, 1989:285-309.

41. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. Journal of Clinical Epidemiology 1992;45:743-760.

42. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. British Medical Journal 1997;314:1580-1583.

43. Gothan A, Brown R, Marsden C. Depression in Parkinson's disease: a quantitative and qualitative analysis. Journal of Neurology, Neurosurgery, and Psychiatry 1986;49:381-389.

44. Brown R, MacCarthy B, Jahanshahi M, Marsden C. Accuracy of self-reported disability in patients with Parkinsonism. Archives of Neurology 1989;46:955-959.

45. Hays R, Vickery B, Hermann B, Perrine K, Cramer J, Meador K, Spritzer K, Devinsky O. Agreement between proxy reports and self-reports of quality of life in epilepsy patients. Quality of Life Research 1995;4:159-165.

46. Dorman P, Waddell F, Slattery J, Dennis M, Sandercock P. Are proxy assessments of health status after stroke with the EuroQoL questionnaire feasible, accurate, and unbiased? Stroke 1997;28:1883-1887.

47. Vickrey BG, Hays RD, Engel J, Spritzer K, Rogers WH, Rausch R, Graber J, Brook RH. Outcome assessment for epilepsy surgery: the impact

of measuring health-related quality of life. Annals of Neurology 1995;37:158-166.

48. Devinsky O. Outcomes research in neurology: incorporating health-related quality of life. Annals of Neurology 1995;37:141-142.

49. Ware JE Jr. Measuring patients' views: the optimum outcome measure. British Medical Journal 1993;306:1429-1430.

50. Bayley KB, London MR, Grunkemeier GL, Lansky DJ. Measuring the success of treatment in patient terms. Medical Care 1995;33:AS226-AS235.

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

52. Gill T, Feinstein A. A critical appraisal of the quality of quality-of-life measurements. Journal of the American Medical Association 1994;272:619-626.

53. Editorial. Quality of life and clinical trials. Lancet 1995;346:1-2.

54. Hunt SM. The problem of quality of life. Quality of Life Research 1997;6:205-212.

55. Muldoon MF, Barger SD, Flory JD, Manuck SB. What are quality of life measurements measuring? British Medical Journal 1998;316:542-545.

56. Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey manual and interpretation guide. Boston, Massachusetts: Nimrod Press, 1993.

57. Hunt SM, McEwan J. The development of a subjective health indicator. Social Health and Illness 1980;2:231-246.

58. EuroQoL Group. EuroQoL: a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.

59. European Study Group on Interferon beta-1b in Secondary Progressive MS. Placebo-controlled multicentre randomised trial of interferon beta-1b in treatment of secondary progressive multiple sclerosis. Lancet 1998;352:1491-1497.

60. Bensimon G, Lacomblez L, Meininger V, Group ARS. A controlled trial of riluzole in amyotrophic lateral sclerosis. New England Journal of Medicine 1994;330:585.

61. Ringel SP, Vickrey BG. Measuring quality of care in neurology. Archives of Neurology 1997;54:1329-1332.

62. Langhorne P, Williams BO, Gilchrist W, Howie K. Do stroke units save lives? Lancet 1993;342:395-398.

63. International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. Lancet 1997;349:1641-1649.

64. Kurtzke JF. A new scale for evaluating disability in multiple sclerosis. Neurology 1955;5:580-583.

65. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. Maryland State Medical Journal 1965;14:61-65.

66. Ashworth B. Preliminary trial of carisoprodol in multiple sclerosis. Practitioner 1964;192:540-542.

67. Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. Neurology 1967;17:427-442.

68. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press, 1992.

69. Miller DH, Albert PS, Barkhof F, Francis G, Frank JA, Hodgkinson S, Lublin FD, Paty DW, Reingold SC, Simon J. Guidelines for the use of

magnetic resonance techniques in monitoring the treatment of multiple sclerosis. Annals of Neurology 1996;39:6-16.

70. Whitaker JN, McFarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis trials: a critical analysis. Multiple Sclerosis 1995;1:37-47.

71. Marsden CD. Editorial: what should neurologists do? Journal of Neurology, Neurosurgery, and Psychiatry 1981;44:1059-1060.

72. Lossef NA, Webb SW, O'Riordan JI, Page R, Wang L, Barker GJ, Tofts PS, McDonald WI, Miller DH, Thompson AJ. Spinal cord atrophy and disability in MS: a new reproducible and sensitive MRI method to monitor disease progression. Brain 1996;119:701-708.

73. Polman CH, Hartung HP. The treatment of multiple sclerosis: current and future. Current Opinion in Neurology 1995;8:200-209.

74. Herndon RM, Murray TJ. Proceedings of the international conference on therapeutic trials in multiple sclerosis. Archives of Neurology 1983;40:663-710.

75. Ellison GW, Myers LW, Leake BD, Mickey MR, Ke D, Syndulko K, Tourtellotte WW. Design strategies for multiple sclerosis clinical trials. Annals of Neurology 1994;1994:S108-S112.

76. Nauta JJP, Thompson AJ, Barkhof F, Miller DH. Magnetic resonance imaging in monitoring the treatment of multiple sclerosis patients: statistical power of parallel-groups and crossover designs. Journal of Neurological Sciences 1994;122:6-14.

77. Foa R. Ethical considerations raised by clinical trials. In: Goodkin DE, Rudick RA, editors. Multiple sclerosis: advances in clinical trial design, treatment and future perspectives. London: Springer-Verlag, 1996:335-350.

78. Weinshenker BG. Natural history of multiple sclerosis. Annals of Neurology 1994;36:S6-S11.

79. Weinshenker BG, Issa M, Baskerville J. Meta-analysis of the placebo-treated groups in clinical trials of progressive MS. Neurology 1996;46:1613-1619.

80. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley, 1986.

81. Cone JD, Foster SL. Training in measurement: always the bridesmaid. American Psychologist 1991;46(6):653-654 (letter).

82. Kurtzke JF. On the evaluation of disability in multiple sclerosis. Neurology 1961;11:686-694.

83. Kurtzke JF, Berlin L. The effects of isoniazid on patients with multiple sclerosis. American Review of Tuberculosis 1954;70:577-591.

84. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 2nd ed. Oxford: Oxford University Press, 1996.

85. Guilford JP. Psychometric methods. 2nd ed. New York: McGraw-Hill, 1954.

86. Nunnally JC Jr. Tests and measurements: assessment and prediction. New York: McGraw-Hill, 1959.

87. Rogers T. The psychological testing enterprise: an introduction. Pacific Grove, California: Brooks / Cole, 1995.

88. Nunnally JC Jr. Introduction to psychological measurement. New York: McGraw-Hill, 1970.

89. Thurstone LL. A method for scaling psychological and educational tests. Journal of Educational Psychology 1925;16:433-451.

90. Thurstone LL. Attitudes can be measured. American Journal of Sociology 1928;33:529-554.

91. Nunnally JC. Psychometric theory. 1st ed. New York: McGraw-Hill, 1967.

92. American Psychological Association. Technical recommendations for educational and psychological tests and diagnostic techniques. Washington, DC: American Psychological Association, 1954.

93. American Educational Research Association, National Council on Measurement used in Education. Technical recommendations for achievement tests. Washington, DC: National Education Association, 1955.

94. American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for educational and psychological tests and manuals. Washington, DC: American Psychological Association, 1966.

95. American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for education and psychological tests. Washington, DC: American Psychological Association, 1974.

96. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Psychological Association, 1985.

97. Ware JE Jr, Brook RH, Davies-Avery A, Williams KN, Stewart AL, Rogers WH, Donald CA, Johnson SA. Conceptualization and measurement of health for adults in the Health Insurance Study. Volume I: Model of health and methodology. Santa Monica, California: The RAND Corporation, 1980 (publication no. R-1987/1-HEW).

98. Brook RH, Ware JE Jr, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, Williams KN, Johnston SA. Overview of adult health status measures fielded in RAND's Health Insurance Study. Medical Care 1979;17(Suppl):1-131.

99. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, McGlynn A, Ware JE Jr. Functional Status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. Journal of the American Medical Association 1989;262:907-913.

100. McHorney CA, Ware JE Jr, Rogers W, Raczek AE, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. Medical Care 1992;30:MS253-MS265.

101. Spitzer WO. State of science 1986: quality of life and functional status as target variables for research. Journal of Chronic Diseases 1987;40:465-471.

102. Kaplan RM, Bush JW, Barry CC. Health status: types of validity and the index of well-being. Health Services Research 1976;11:478-507.

103. Sackett DL, Chambers IW, McPherson AS, Goldsmith CH, McCauley RG. The development and application of indices of health: general methods and a summary of results. American Journal of Public Health 1977;67:423-428.

104. Bombardier C, Tugwell P. A methodological framework to develop and select indices for clinical trials: statistical and judgemental approaches. Journal of Rheumatology 1982;9:753-757.

105. Deyo R. Measuring functional outcomes in therapeutic trials for chronic disease. Controlled Clinical Trials 1984;5:223-240.

106. Ware JE Jr. Methodological considerations in the selection of health status assessment procedures. In: Wenger NK, Mattson ME, Furberg CD, editors. Assessment of quality of life in clinical trials of cardiovascular therapies. New York: Le Jacq, 1984:87-111.

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

108. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. Journal of Chronic Diseases 1986;39:897-906.

109. Lamping DL. Assessment in health psychology. Canadian Psychology 1985;26:121-139.

110. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronic Diseases 1987;40:171-178.

111. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 1st ed. Oxford: Oxford University Press, 1987.

112. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. Journal of Clinical Epidemiology 1989;42:403-408.

113. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 1st ed. Oxford: Oxford University Press, 1989.

114. Bowling A. Measuring health: a review of quality of life measurement scales. Milton Keynes: Open University Press, 1991.

115. Task Force on Standards for Measurement in Physical Therapy. Standards for tests and measurements in physical therapy practice. Physical Therapy 1991;71:589-622.

116. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Controlled Clinical Trials 1991;12:142s-158s.

117. Guyatt GH, Krishner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? Journal of Clinical Epidemiology 1992;45:1341-1345.

118. Johnston MV, Keith RA, Hinderer SR. Measurement standards for interdisciplinary medical rehabilitation. Archives of Physical Medicine and Rehabilitation 1992;73:S3-S23.

119. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I: Applications and uses in assessment. British Medical Journal 1992;305:1074-1077.

120. Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: Design, analysis, and interpretation. British Medical Journal 1992;305:1145-1148.

121. Spiegelhalter D, Gore S, Fitzpatrick R, Fletcher A, Jones D, Cox D. Quality of life measures in health care. III: Resource allocation. British Medical Journal 1992;305:1205-1208.

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

123. Guyatt GH, Freeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine 1993;118:622-629.

124. Ware JE Jr. Standards for evaluating the standards. Medical Outcomes Trust Bulletin 1994;2(3):2-3.

125. Bowling A. Measuring disease: a review of disease-specific quality of life measurement scales. Buckingham: Open University Press, 1995.

126. Hobart JC, Lamping DL, Thompson AJ. Evaluating neurological outcome measures: the bare essentials. Journal of Neurology, Neurosurgery, and Psychiatry 1996;60:127-130.

127. McDowell I, Jenkinson C. Development standards for health measures. Journal of Health Services Research and Policy 1996;1:238-246.

128. Williams JI, Naylor CD. How should health status instruments be assessed? Cautionary notes on procrustean frameworks. Journal of Clinical Epidemiology 1992;45:1347-1351.

129. Tarlov AR. Scientific Advisory Committee appointed. Medical Outcomes Trust Bulletin 1994;2(2):1.

130. Scientific Advisory Committee of the Medical Outcomes Trust. Instrument review criteria. Medical Outcomes Trust Bulletin 1995;3(4):I-IV.

131. Mackenzie CR, Charleston ME, DiGioia D, Kelley K. A patient-specific measure of change in maximal function. Archives of Internal Medicine 1986;146:1325-1329.

132. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Medical Care 1989;27(Suppl):S178-S189.

133. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis and Rheumatism 1985;28:542-547.

134. Norman GR. Issues in the use of change scores in randomized trials. Journal of Clinical Epidemiology 1989;42:1097-1105.

135. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopaedic evaluation. Medical Care 1990;28:632-638.

136. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. Journal of Clinical Epidemiology 1991;44:417-421.

137. Juniper E, Guyatt G, Goldstein R. Determining a minimal important change in a disease-specific quality of life instrument. Journal of Clinical Epidemiology 1994;47:81-87.

138. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. Journal of Clinical Epidemiology 1997;50:869-879.

139. Cronbach LJ, Furby L. How we should measure "change" - or should we? Psychological Bulletin 1970;74:68-80.

140. Nunnally JC. The study of change in evaluation research: principles concerning measurement, experimental design, and analysis. In: Struening E, Guttentag M, editors. Handbook of evaluation research. Beverley Hills, California: Sage, 1975:101-137.

141. National Audit Office. Physical disability 1986 and beyond. A report by the controller and auditor general. London: Royal College of Physicians, 1987.

142. Neurological Charities. Neurological provision: key areas and targets, response to the "Health of the Nation" by the neurological charities. London: Department of Health, 1991.

143. Association of British Neurologists, British Society for Rehabilitation Medicine. Neurological rehabilitation in the UK: report of a working party. London: Association of British Neurologists, 1992.

144. Greenwood RJ, Barnes M, McLellan DL, editors. Neurological rehabilitation. Edinburgh: Churchill Livingstone, 1993.

145. Scheinberg LC. Therapeutic strategies. Annals of Neurology 1994;36:S122-S129.

146. Thompson AJ. Rehabilitation of progressive neurological disorders: a worthwhile challenge ? Current Opinion in Neurology 1996;9:437-440.

147. La Rocca NG, Shapiro RT, Scheinberg LC, Kraft GH. Comprehensive care in multiple sclerosis: the whole versus the parts. Journal of Neurological Rehabilitation 1994;8:95-98.

148. McGrath JR, Davis AM. Rehabilitation: where do we go and how do we get there? Clinical Rehabilitation 1992;6:225-235.

149. Wade DT. Measurement in neurological rehabilitation. Current Opinion in Neurology 1993;6:778-784.

150. Freeman JA. The efficacy of inpatient rehabilitation in multiple sclerosis. PhD thesis, University of London, 1997.

151. Langton-Hewer R. Neurorehabilitation in the UK. International Journal of Rehabilitation Medicine 1980;2:116-125.

152. Greenwood RJ. Neurology and rehabilitation in the UK: a view. Journal of Neurology, Neurosurgery, and Psychiatry 1992;55:51-53.

153. Edwards SM. Longer-term management for patients with residual or progressive disability. In: Edwards SM, editor. Neurological physiotherapy: a problem-solving approach. London: Churchill Livingstone, 1996:189-206.

154. Dobkin B, Thompson AJ. The principles of neurological rehabilitation. In: Bradley WG, Daroff RB, Fenichel GM, Marsden CD, editors. Neurology in clinical practice: principles of diagnosis and management. Boston, Massachusetts: Butterworth-Heinmann, in press.

155. World Health Organization. International Classification of Impairments, Disabilities and Handicaps (ICIDH): a manual of classification relating to the consequences of disease. Geneva: World Health Organization, 1980.

156. Chamie M. The status and use of the International Classification of Impairments, Disabilities, and Handicaps (ICIDH). World Health Statistics Quarterly 1990;43:273-280.

157. Wood PHN. Appreciating the consequences of disease: the International Classification of Impairments, Disabilities, and Handicaps. WHO Chronicle 1980;34:376-380.

158. Robinson D. The International Classification of Impairments, Disabilities, and Handicaps. International Rehabilitation Medicine 1985;7:60-67.

159. World Health Organization. ICIDH-2: International classification of impairments, activities, and participation. A manual of dimensions of disablement and functioning. Beta-1 draft for field trials. Geneva: World Health Organisation, 1997.

160. Bickenbach JE, Chatterji S, Badley EM, Ustin TB. Models of disablement, universalism and the International Classification of Impairments, Disabilities, and Handicaps. Social Science and Medicine 1999;48:1173-1187.

161. Gresham GE, Labi MLC. Functional assessment instruments currently available for documenting outcomes in rehabilitation medicine. In: Granger CV, Gresham GE, editors. Functional assessment in rehabilitation medicine. Baltimore, Maryland: Williams and Wilkins, 1984:65-85.

162. Rubenstein LZ, Schairer C, Wieland GD, Kane R. Systematic biases in functional status assessment of elderly adults: effects of different data sources. Journal of Gerontology 1984;39:686-691.

163. Granger CV, Hamilton BB, Keith RA, Zielezny M, Sherwin FS. Advances in functional assessment for medical rehabilitation. Topics in Geriatric Rehabilitation 1986;1(3):59-74.

164. Keith RA. Functional assessment measures in medical rehabilitation: current status. Archives of Physical Medicine and Rehabilitation 1984; 65:74-78.

165. Hamilton BB, Granger CV, Sherwin FS, Zielezny M, Tashman JS. A uniform national data system for medical rehabilitation. In: Fuhrer MJ, editor. Rehabilitation outcomes: analysis and measurement. Baltimore, Maryland: Paul H Brookes, 1987:137-147.

166. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The Functional Independence Measure: a new tool for rehabilitation. In: Eisenberg MG,

Grzesiak RC, editors. Advances in clinical rehabilitation. New York: Springer-Verlag, 1987:6-18.

167. Feinstein AR, Josephy BR, Wells CK. Scientific and clinical problems in indexes of functional disability. Annals of Internal Medicine 1986;105:413-420.

168. Herndon RM, editor. Handbook of neurological rating scales. New York: Demos, 1997.

169. Frey WD. Functional assessment in the '80's: a conceptual enigma, a technical challenge. In: Halpern AS, Fuhrer MJ, editors. Functional assessment in rehabilitation. Baltimore, Maryland: Paul H Brookes, 1984:11-43.

170. Deyo R, Patrick D. Barriers to the use of health status measures in clinical investigation, patient care and policy research. Medical Care 1989;27(Suppl):S254-S268.

171. Slater SB, Vukmanovic C, Macukanovic P, Prulovic T, Cutler JL. The definition and measurement of disability. Social Science and Medicine 1974;8:305-308.

172. Krauze EA. The political sociology of rehabilitation. In: Albrecht GL, editor. The sociology of physical disability and rehabilitation. Pittsburgh: University of Pittsburgh Press, 1976.

173. Duckworth D. The classification and measurement of disablement. London: Department of Health and Social Security, Social Research Branch, Research report no. 10, 1983.

174. Wood PHN, Badley EM. People with disabilities. New York: World Rehabilitation Fund, 1981.

175. Martin J, Meltzer M, Elliot D. OPCS surveys of disability in Great Britain. Report 1: the prevalence of disability among adults. London: Her Majesty's Stationary Office, 1988.

176. Kempen GIJM, Meedema I, Ormel J, Molenaar W. The assessment of disability with the Groningen Activity Restriction Scale: conceptual framework and psychometric properties. Social Science and Medicine 1996;43:1601-1610.

177. Nagi SZ. The concept and measurement of disability. In: Berkowitz ED, editor. Disability policies and government programs. New York: Preager Press, 1979:1-15.

178. Nagi SZ. Disability concepts revisited: implications for prevention. In: Pope AM, Tarlov AR, editors. Disability in America: toward a national standard for prevention. Washington, DC: National Academy Press, 1991:309-339.

179. Badley EM. The ICIDH: format, application in different settings, and distinction between disability and handicap. International Disability Studies 1987;9:122-125.

180. Bury MR. The ICIDH: a review of research and prospects. International Disability Studies 1987;9:118-128.

181. American Medical Association. Guide to the evaluation of permanent impairment. 2nd ed. Chicago: American Medical Association, 1984.

182. Mooney V. Impairment, disability, and handicap. Clinical Orthopaedics 1987;221:14-25.

183. Luck JV, Florence DW. A brief history and comparative analysis of disability systems and impairment rating guides. Orthopaedic Clinics of North America 1988;19:839-844.

184. Sheldon MP. A physical achievement record for use with crippled children. Journal of Health and Physical Education 1935;6:30-34.

185. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. Journal of the American Medical Association 1963;185:914-919.

186. Bombardier C, Tugwell P. Methodological considerations in functional assessment. Journal of Rheumatology 1987;14:6-10.

187. Alexander JL, Furher MJ. Functional assessment in individuals with physical impairments. In: Halpern AS, Furher MJ, editors. Functional assessment in rehabilitation. Baltimore, Maryland: Paul H Brookes, 1984:45-49.

188. Van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJA, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 1988;19:604-607.

189. Bamford J, Sandercock P, Dennis M, Slattery J, Warlow CP. A prospective study of acute cerebrovascular disease in the community: the Oxfordshire Community Stroke Project 1981-1986. I. Methodology, demography and incidence of cases of first-ever stroke. Journal of Neurology, Neurosurgery, and Psychiatry 1988;51:1373-1380.

190. Barnes MP. FIM and FAM - a positive workshop. European Rehabilitation Newsletter 1995;7(February):3 and 7.

191. Hall KM, Hamilton BB, Gordon WA, Zasler ND. Characteristics and comparisons of functional assessment indices: Disability Rating Scale, Functional Independence Measure and Functional Assessment Measure. Journal of Head Trauma and Rehabilitation 1993;8:60-74.

192. Albrecht GL, Harasymiw SJ. Evaluating rehabilitation outcomes by cost function indicators. Journal of Chronic Diseases 1979;32:525-533.

193. Granger CV. Outcome of comprehensive medical rehabilitation: an analysis based upon the impairment, disability, and handicap model. International Rehabilitation Medicine 1985;7:45-50.

194. Hamilton BB, Laughlin JA, Granger CV, Kayton RM. Interrater agreement of the seven level Functional Independence Measure (FIM). Archives of Physical Medicine and Rehabilitation 1991;72:720 (abstract).

195. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the Functional Independence Measurement and its performance among rehabilitation inpatients. Archives of Physical Medicine and Rehabilitation 1993;74:531-536.

196. Granger CV. Guide for the Uniform Data Set for Medical Rehabilitation (Adult FIM) Version 4.0. Buffalo, New York: UB Foundation Activities, Inc., 1993.

197. Marolf MV, Vaney C, Konig N, Schenk T, Prosiegel M. Evaluation of disability in multiple sclerosis patients: a comparative study of the Functional Independence Measure, the Extended Barthel Index and the Expanded Disability Status Scale. Clinical Rehabilitation 1996;10:309-313.

198. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

199. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

200. Segal ME, Ditunno JF, Staas WE. Interinstitutional agreement of individual Functional Independence Measure (FIM) items measured at two sites on one sample of SCI patients. Paraplegia 1993;31:622-631.

201. Brosseau L. The inter-rater reliability and construct validity of the Functional Independence Measure for multiple sclerosis subjects. Clinical Rehabilitation 1994;8:107-115.

202. Hamilton BB, Laughlin JA, Fielder RC, Granger CV. Interrater reliability of the 7-level Functional Independence Measure (FIM). Scandinavian Journal of Rehabilitation Medicine 1994;26:115-119.

203. Chau N, Daler S, Andre JM, Patris A. The inter-rater agreement of two functional independence scales: the Functional Independence Measure and a subjective uniform continuous scale. Disability and Rehabilitation 1994;16(2):63-71.

204. Ottenbacher KJ, Hsu Y, Granger CV, Fielder RC. The reliability of the Functional Independence Measure: a quantitative review. Archives of Physical Medicine and Rehabilitation 1996;77:1226-1232.

205. Stineman MG, Shea JA, Jette A, Tassoni CJ, Ottenbach KJ, Fielder R, Granger CV. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. Archives of Physical Medicine and Rehabilitation 1996;77:1101-1108.

206. Stineman MG, Jette A, Fiedler R, Granger CV. Impairment-specific dimensions within the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 1997;78:636-43.

207. Sharrack B, Hughes RAC, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999;122:141-159.

208. Granger CV, Cotte ACR, Hamilton BB, Fielder RC, Hens MM. Functional assessment scales: a study of persons with multiple sclerosis. Archives of Physical Medicine and Rehabilitation 1990;71:870-875.

209. Davidoff GN, Roth EJ, Haughton JS, Ardner MS. Cognitive dysfunction in spinal cord injury patients: sensitivity of the Functional Independence Measure subscales vs. neuropsychologic assessment. Archives of Physical Medicine and Rehabilitation 1990;71:326-329.

210. Disler PB, Roy CW, Smith BP. Predicting hours of care needed. Archives of Physical Medicine and Rehabilitation 1993;74:139-143.

211. Granger CV, Cotter AC, Hamilton BB, Fielder RC. Functional assessment scales: a study of persons after stroke. Archives of Physical Medicine and Rehabilitation 1993;74:133-138.

212. Kaplan CP, Corrigan JD. The relationship between cognition and functional independence in adults with traumatic brain injury. Archives of Physical Medicine and Rehabilitation 1994;75:643-647.

213. van der Putten JJMF, Hobart JC, Freeman JA, Thompson AJ. Measuring change in rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. Journal of Neurology, Neurosurgery, and Psychiatry 1999;66:480-484.

214. McPherson KM, Pentland B, Cudmore SF, Prescott RJ. An inter-rater reliability study of the Functional Assessment Measure (FIM+FAM). Disability and Rehabilitation 1996;18:341-347.

215. Alcott D, Dixon K, Swann R. The reliability of the items of the Functional Assessment Measure (FAM): differences in abstractness between FAM items. Disability and Rehabilitation 1997;19:355-358.

216. McPherson KM, Pentland B. Disability in patients following traumatic brain injury - which measure? International Journal of Rehabilitation Research 1997;20:1-10.

217. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill, 1978.

218. Patrick D, Deyo R. Generic and disease-specific measures in assessing health status and quality of life. Medical Care 1989;27(Suppl):S217-S232.

219. Ware JE Jr. Standards for validating health measures: definition and content. Journal of Chronic Diseases 1987;40:473-480.

220. Wylie CM, White BK. A measure of disability. Archives of Environmental Health 1964;8:834-839.

221. Shah S, Vanclay F, Cooper B. Improving the sensitivity of the Barthel Index for stroke rehabilitation. Journal of Clinical Epidemiology 1989;42:703-709.

222. Gresham GE, Phillips TF, Labi LC. ADL status in stroke: relative merits of three standard indexes. Archives of Physical Medicine and Rehabilitation 1980;61:355-538.

223. Shinar D, Gross C, Bronstein K, Licata-Gehr E, Eden D, Cabrera A, Fishman I, Roth A, Barwick J, Kunitz S. Reliability of the Activities of Daily Living Scale and its use in telephone interview. Archives of Physical Medicine and Rehabilitation 1987;68:723-728.

224. Loewen SC, Anderson BA. Reliability of the Modified Motor Assessment Scale and the Barthel Index. Physical Therapy 1988;68:1077-1081.

225. Barer DH, Murphy JJ. Scaling the Barthel: a 10-point hierarchical version of the activities of daily living index for use with stroke patients. Clinical Rehabilitation 1993;7:271-277.

226. Wade DT, Collin C. The Barthel ADL Index: a standard measure of disability? International Disability Studies 1988;10:64-67.

227. Royal College of Physicians. Standardised assessment scales for elderly people: report of joint workshops of the Research Unit of the Royal College of Physicians and the British Geriatrics Society. London: Royal College of Physicians of London, 1992.

228. Granger CV, Albrecht GL, Hamilton BB. Outcome of comprehensive medical rehabilitation: measurement by PULSES Profile and Barthel Index. Archives of Physical Medicine and Rehabilitation 1979;60:145-154.

229. Roy CW, Togneri J, Hay E, Pentland B. An inter-rater reliability study of the Barthel Index. International Journal of Rehabilitation Research 1988;11:67-70.

230. Wolfe CDA, Taub NA, Woodrow EJ, Burney PGJ. Assessment of scales of disability and handicap for stroke patients. Stroke 1991;22:1242-1244.

231. Gompertz P, Pound P, Ebrahim S. The reliability of stroke outcome measures. Clinical Rehabilitation 1993;7:290-296.

232. Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. Clinical Rehabilitation 1994;8:233-239.

233. Wade DT, Langton Hewer R. Functional abilities after stroke: measurement, natural history and prognosis. Journal of Neurology, Neurosurgery and Psychiatry 1987;50:177-182.

234. McPherson K, Sloan RL, Hunter J, Dowell CM. Validation studies of the OPCS scale - more useful than the Barthel Index? Clinical Rehabilitation 1993;7:105-112.

235. Gompertz P, Pound P, Ebrahim S. Validity of the Extended Activities of Daily Living Scale. Clinical Rehabilitation 1994;8:275-280.

236. van Bennekom CAM, Jelles F, Lankhorst GJ. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. Journal of Clinical Epidemiology 1996;49:39-44.

237. Novak S, Johnson J, Greenwood R. Barthel revisited: making guidelines work. Clinical Rehabilitation 1996;10:128-134.

238. Goodkin DE, Cookfair D, Wende K, Bourdette D, Pullicino P, Scherokman B, Whitman R. Inter- and intra-rater scoring agreement using grades 1.0 to 3.5 of the Kurtzke Expanded Disability Status Scale (EDSS). Neurology 1992;42:859-863.

239. Willoughby EW, Paty DW. Scales for rating impairment in multiple sclerosis: a critique. Neurology 1988;38:1793-1798.

240. Noseworthy JH, Vander voort MK, Wong CJ, Ebers GC. Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. Neurology 1990;40:971-975.

241. Harwood RH, Ebrahim S. Manual of the London Handicap Scale. Nottingham: Department of Health Care of the Elderly, University of Nottingham, 1995.

242. Medical Outcomes Trust. Involving physicians in health outcomes assessment. Medical Outcomes Trust Bulletin 1996;4(2):1.

243. Ware JE Jr, Kosinski MA, Keller SD. SF-36 physical and mental health summary scales: a user's manual. Boston, Massachusetts: The Health Institute, New England Medical Centre, 1994.

244. Ware JE Jr, Sherbourne DC. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. Medical Care 1992;30:473-483.

245. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care 1993;31:247-263.

246. McHorney CA, Ware JE Jr, Lu JFR, Sherbourne CD. 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 1994;32:40-66.

247. Stewart AL, Hays RD, Ware JE Jr. The MOS Short-Form General Health Survey: reliability and validity in a patient population. Medical Care 1988;26:724-735.

248. Ware JE Jr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Medical Care 1995;33:AS264-AS279.

249. Goldberg DP. Manual of the General Health Questionnaire. Windsor: NFER-Nelson, 1978.

250. Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. Psychological Medicine 1979;9:139-145.

251. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:189-198.

252. Wechsler D. The Wechsler Adult Intelligence Test-Revised. New York: The Psychological Corporation, 1981.

253. Langdon DW. Neuropsychological problems and solutions. In: Edwards SM, editor. Neurological physiotherapy. London: Churchill Livingstone, 1996:41-61.

254. DeFilippis NA, McCampbell E. The Halstead Booklet Category Test. Odessa, Florida: Psychological Assessment Resources, 1993.

255. Grant DA, Berg EA. Wisconsin Card Sorting Test. Odessa, Florida: Psychological Assessment Resources, 1993.

256. Langdon DW, Warrington EK. Verbal and Spatial Reasoning Test. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1995.

257. Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test. San Antonio, Texas: Psychological Corporation, 1987.

258. Lezak MD. Neuropsychological assessment. 3rd ed. Oxford: Oxford University Press, 1995.

259. Warrington EK. Recognition Memory Test. Windsor: Nelson, 1984.

260. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Transition questions to assess outcomes in rheumatoid arthritis. British Journal of Rheumatology 1993;32:807-811.

261. Ware JE Jr, Davies-Avery A, Donald CA. Conceptualization and measurement of health for adults in the Health Insurance Study. Volume V: General health perceptions. Santa Monica, California: The RAND Corporation, 1978 (publication no. R-1987/5-HEW).

262. Essink-Bot M-L, Krabbe PFM, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures: the Nottingham Health Profile, the Medical Outcomes Study 36-Item Short-Form Health Survey, the COOP/WONCA charts and the EuroQol instrument. Medical Care 1997;35:522-537.

263. Holmes WC, Bix B, Shea JA. SF-20 score and item distributions in a human immunodeficiency virus-seropositive sample. Medical Care 1996;34:562-569.

264. Leplege A, Rude N, Ecosse E, Ceinos R, Dohin E, Pouchot J. Measuring quality of life from the point of view of HIV-positive subjects: the HIV-QL31. Quality of Life Research 1997;6:585-594.

265. Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. Medical Care 1979;17:902-921.

266. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Quality of Life Research 1995;4:293-307.

267. Holmes WC, Shea JA. Performance of a new, HIV/AIDS-targeted quality of life (HAT-QoL) instrument in asymptomatic seropositive individuals. Quality of Life Research 1997;6:561-571.

268. Anastasi A, Urbina S. Psychological testing. 7th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1997.

269. Spector PE. Summated rating scale construction: an introduction. Newbury Park, California: Sage, 1992.

270. Zubin J. The method of internal consistency for selecting items. Journal of Educational Psychology 1934;25:345-356.

271. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill, 1994.

272. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.

273. Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. Journal of Clinical Epidemiology 1991;44:381-390.

274. Kristof W. The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika 1963;28:221-238.

275. Fiske DW. Some hypotheses concerning test adequacy. Educational and Psychological Measurement 1966;26:69-88.

276. Tyler TA, Fiske DW. Homogeneity indices and test length. Educational and Psychological Measurement 1968;28:767-777.

277. Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology 1993;78:98-104.

278. Cronbach LJ. Essentials of psychological testing. 5th ed. New York: Harper Collins, 1990.

279. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychological Reports 1966;19:3-11.

280. Bartko JJ, Carpenter WT. On the methods and theory of reliability. Journal of Nervous and Mental Disease 1976;163:307-317.

281. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-428.

282. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1996;1:30-46.

283. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin 1955;52:281-302.

284. Cronbach LJ. Essentials of psychological testing. New York: Harper and Row, 1949.

285. Gulliksen H. Theory of mental tests. New York: Wiley, 1950.

286. Green SB, Lissitz RW, Muliak SA. Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement 1977;37:827-838.

287. Kerlinger FN. Foundations of behavioural research. 2nd ed. New York: Holt, Rinehart, and Winston, 1973.

288. Cohen J. The earth is round (p < .05). American Psychologist 1994;49:997-1003.

289. Cortina JM, Dunlap WP. On the logic and purpose of significance testing. Psychological Methods 1997;2:161-172.

290. Hays R, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Quality of Life Research 1992;1:73-73.

291. Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Physical Therapy 1996;76:1109-1123.

292. Liang MH. Evaluating instrument responsiveness. Journal of Rheumatology 1995;22:1191-1192.

293. Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

294. Sechrest L. Incremental validity. In: Jackson D, Messick S, editors. Problems in human assessment. New York: McGraw-Hill, 1967:368-371.

295. Bradley WG, Daroff RB, Fenichel GM, Marsden CD, editors. Neurology in clinical practice: principles of diagnosis and management. 2nd ed. Boston, Massachusetts: Butterworth-Heinemann, 1997.

296. Likert RA. A technique for the development of attitudes. Archives of Psychology 1932;140:5-55.

297. Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of life questionnaire construction. Journal of Clinical Epidemiology 1997;50:233-238.

298. Hattie J. Methodological review: assessing unidimensionality of tests and items. Applied Psychological Measurement 1985;9:139-164.

299. Brook RH, Ware JE Jr, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, Williams KN, Johnson SA. Conceptualization and measurement of health for adults in the Health Insurance Study. Volume VIII: Overview. Santa Monica, California: The RAND Corporation, 1979 (publication no. R-1987/8-HEW).

300. Likert RA, Roslow S, Murphy G. A simple and reliable method of scoring the Thurstone attitude scales. Journal of Social Psychology 1934;5:228-238.

301. Edwards AL. Techniques of attitude scale construction. New York: Appleton-Century-Crofts, 1957.

302. Ware JE Jr, Harris WJ, Gandek B, Rogers BW, Reese PR. MAP-R for windows: multitrait / multi-item analysis program - revised user's guide. Boston, MA: Health Assessment Lab., 1997.

303. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 1959;56:81-105.

304. Hays RD, Hayashi T. Beyond internal consistency reliability: rationale and user's guide for Multi-Trait Analysis Program on the microcomputer. Behavior Research Methods, Instruments, & Computers 1990;22:167-175.

305. Fayers PM, Machin D. Factor analysis. In: Staquet MJ, Hays RD, Fayers PM, editors. Quality of life assessment in clinical trials: methods and practice. Oxford: Oxford University Press, 1998:191-223.

306. Guttman LA. Some necessary conditions for common-factor analysis. Psychometrika 1954;19:149-161.

307. Cattell RB. The scree test for the number of factors. Multivariate Behavioural Research 1966;1:245-276.

308. Kline P. An easy guide to factor analysis. London: Routledge, 1994.

309. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Quality of Life Research 1997;6:139-150.

310. Hays RD, Anderson R, Revicki DA. Psychometric considerations in evaluating health-related quality of life measures. Quality of Life Research 1993;2:441-449.

311. Cohen J. Statistical power analysis for the behavioural sciences. 1st ed. Hillside, New Jersey: Lawrence Erlbaum, 1969.

312. Wright JG, Young NL. A comparison of different indices of responsiveness. Journal of Clinical Epidemiology 1997;50:239-246.

313. Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: British Medical Journal Books, 1998.

314. Brosseau L, Phillippe P, Potvin L, Boulanger Y-L. Post-stroke inpatient rehabilitation: I. Predicting length of stay. American Journal of Physical Medicine and Rehabilitation 1996;75:422-430.

315. Brosseau L, Potvin L, Phillippe P, Boulanger Y-L. Post-stroke inpatient rehabilitation: II. Predicting discharge disposition. American Journal of Physical Medicine and Rehabilitation 1996;75:431-436.

316. Nunnally JC. Introduction to statistics for psychology and education. New York: McGraw-Hill, 1975.

317. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Archives of Physical Medicine and Rehabilitation 1994;75:127-132.

318. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. Journal of Clinical Epidemiology 1995;48:1369-1378.

319. Nyein K, McMichael L, Turner-Stokes L. Can a Barthel Index score be derived from the FIM? Clinical Rehabilitation 1999;13:56-63.

320. McIver JP, Carmines EG. Unidimensional scaling. Newbury Park, California: Sage, 1981.

321. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980.

322. Kiresuk TJ, Smith A, Cardillo JE, editors. Goal Attainment Scaling: applications, theory, and measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1994.

323. Freeman JA, Langdon DW, Hobart JC, Thompson AJ. The impact of inpatient rehabilitation on progressive multiple sclerosis. Annals of Neurology 1997;42:236-244.

324. Smith A, Cardillo JE, Smith SC, Amezaga AM. Improvement Scaling (rehabilitation version): a new approach to measuring progress of patients in achieving their individual rehabilitation goals. Medical Care 1998;36:333-347.

325. Jette AM, Davies AR, Cleary PD. The Functional Status Questionnaire: reliability and validity when used in primary care. Journal of General Internal Medicine 1986;1:143-149.

326. Nelson E, Wasson J, Kirk J. Assessment of function in routine clinical practice: description of the COOP chart method and preliminary findings. Journal of Chronic Diseases 1987;40 (Suppl):55S-63S.

327. Parkerson GR, Ghelbach SH, Wagner EH, James SA, Clapp NE, Muhlbaier LH. The Duke-UNC health profile: an adult health status instrument for primary care. Medical Care 1981;19:806-828.

328. Bakheit AMO, Harries SR, Hull RG. Validity of a self-administered version of the Barthel Index in patients with rheumatoid arthritis. Clinical Rehabilitation 1995;9:234-237.

329. Hobart JC, Lamping DL, Thompson AJ. Measuring disability in neurological disease: validity of the self-report Barthel Index. Journal of Neurology 1996;243(Suppl 2):S25 (abstract).

330. Freeman JA, Langdon DW, Hobart JC, Thompson AJ. Inpatient rehabilitation in multiple sclerosis: do the benefits carry over into the community ? Neurology 1999;52:50-56.

331. Warlow CP, Dennis MS, van Gijn J, Hankey GJ, Sandercock PAG, Bamford JM, Wardlaw J. Stroke: a practical guide to management. Oxford: Blackwell Science, 1996.

332. Collin C, Wade DT, Davis S, Horne V. The Barthel ADL Index: a reliability study. International Disability Studies 1988;10:61-63.

333. Bowers DN, Kofroth LK. Comparison of Disability Rating Scale and Functional Independence Measure during recovery from traumatic brain injury. Archives of Physical Medicine and Rehabilitation 1989;70:A-58 (abstract).

334. Dawson J, Fitzpatrick R, Murray D, Carr A. Comparison of measures to assess outcomes in total hip replacement surgery. Quality in Health Care 1996;5:81-88.

335. Koziol JA, Frutos A, Sipe JC, Romine JS, Beutler E. A comparison of two neurologic scoring instruments for multiple sclerosis. Journal of Neurology 1996;243:209-213.

336. Langfitt JT. Comparison of the psychometric characteristics of three quality of life measures in intractable epilepsy. Quality of Life Research 1995;4:101-114.

337. Pinholt EM, Kroenke K, Hanley JF, Kussman MJ, Twyman PL, Carpenter JL. Functional assessment of the elderly: a comparison of standard instruments with clinical judgement. Archives of Internal Medicine 1987;147:484-488.

338. Prieto L, Alonso J, Ferrer M, Anto J. Are results of the SF-36 Health Survey and the Nottingham Health Profile similar? A comparison on COPD patients. Journal of Clinical Epidemiology 1997;50:463-473.

339. Vickrey B, Hays R, Genovese B, Myers L, Ellison G. Comparison of a generic to disease-targeted health-related quality-of-life measures for multiple sclerosis. Journal of Clinical Epidemiology 1997;50:557-569.

340. Jette AM. Functional capacity evaluation: an empirical approach. Archives of Physical Medicine and Rehabilitation 1980;61:85-89.

341. Stewart AL, Ware JE Jr, Brook RH, Davies-Avery A. Conceptualization and measurement of health for adults in the Health Insurance Study. Volume II: Physical health in terms of functioning. Santa Monica, California: The RAND Corporation, 1978 (publication no. R-1987/2-Hew).

342. Thomas VS, Rockwood K, McDowell I. Multidimensionality in instrumental and basic activities of daily living. Journal of Clinical Epidemiology 1998;51:315-321.

343. Beaton D, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. Journal of Clinical Epidemiology 1997;50:79-93.

344. Bessette L, Sangha O, Kuntz KM, Keller B, Lew RA, Fossel AH, Katz JN. Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Medical Care 1998;36:491-502.

345. Murawski MM, Miederhoff PA. On the generalisability of statistical expressions of health related quality of life instrument responsiveness: a data synthesis. Quality of Life Research 1998;7:11-22.

346. Komorita SS, Graham WK. Number of scale points and the reliability of scales. Educational and Psychological Measurement 1965;25:987-995.

347. Masters JR. The relationship between number of response categories and reliability of Likert-type questionnaires. Journal of Educational Measurement 1974;11:49-53.

348. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically significant change. Controlled Clinical Trials 1989;10:407-415.

349. Cronbach LJ. The two disciplines of scientific psychology. American Psychologist 1957;12:671-684.

350. Long WB, Sacco WJ, Coombes SS, Copes WS, Bullock A, Melville JK. Determining normative standards for Functional Independence Measure transitions in rehabilitation. Archives of Physical Medicine and Rehabilitation 1994;75:144-148.

351. Ware JE Jr. Content-based interpretation of health status scores. Medical Outcomes Trust Bulletin 1994;2(4):3.

352. Lydick E, Epstein RS. Interpretation of quality of life changes. Quality of Life Research 1993;2:221-226.

353. Agree EM. The influence of personal care and assistive devices on the measurement of disability. Social Science and Medicine 1999;48:427-443.

354. Ware JE Jr, Manning WGJ, Duan N, Wells KB, Newhouse JP. Health status and the use of outpatient mental health services. American Psychologist 1984;39:1090-1100.

355. Testa MA, Anderson RB, Nackley JF, Hollenberg NK, and the Quality of Life Hypertension Study Group. Quality of life and hypertensive therapy in men: a comparison of captopril with enalapril. New England Journal of Medicine 1993;328:907-913.

356. Hadorn DC, Uerbersax J. Large scale health outcomes evaluation: how should quality of life be measured? Part 1 - Calibration of a brief questionnaire and a search for preference subgroups. Journal of Clinical Epidemiology 1995;48:607-618.

357. Brook RH, Ware JE Jr, Rogers WH, Keeler EB, Davies AR, Donald CA. Does free care improve adult health? Results from a randomised controlled trial. New England Journal of Medicine 1983;309:1426-1434.

358. Cornwall A. Body mapping in health. Rapid Rural Appraisal Series 1991;12:69-76.

359. Deyo RA, Andersson G, Bombardier C, Cherkin DC, Keller RB, Lee CK. Outcome measures for studying patients with low back pain. Spine 1994;18 (Suppl):2032S-2036S.

360. Lydick E, Yawn BP. Clinical interpretation of health-related quality of life data. In: Staquet MJ, Hays RD, Fayers PM, editors. Quality of life assessment in clinical trials: methods and practice. Oxford: Oxford University Press, 1998:299-314

361. Thompson AJ, Colville PL, Ketelaer P, Paty DW. Long term management of multiple sclerosis. MS Management 1994;1:1-9.

362. Badley EM. An introduction to the concepts and classifications of the International Classification of Impairments, Disabilities, and Handicaps. Disability and Rehabilitation. 1993;15:161-178.

363. Pearce G, Kirshner L. Handicap: the focus of multiple sclerosis rehabilitation. MS Management 1995;2(1):21-25.

364. Wade DT. Outcome measurement and rehabilitation (editorial). Clinical Rehabilitation 1999;13:93-95.

365. Eddy DM. Should we change the rules for evaluating medical technologies? In: Gelijns AC, editor. Modern methods of clinical investigation. Washington, DC: National Academic Press, 1990:117-135

366. Wilson BA. How do we know that rehabilitation works ? Neuropsychological Rehabilitation 1993;3:1-4.

367. Black N. Why we need observational studies to evaluate the effectiveness of health care. British Medical Journal 1996;312:1215-1218.

368. Kiresuk TJ, Sherman RE. Goal Attainment Scaling: a general method for evaluating comprehensive community health programs. Community Mental Health Journal 1968;4:443-453.

369. Browne JP, O'Boyle CA, McGee HM, McDonald NJ, Joyce CRB. Development of a direct weighting procedure for quality of life domains. Quality of Life Research 1997;6:301-309.

370. Marx RG, Bombardier C, Hogg-Johnson S, Wright JG. Clinimetric and psychometric strategies for development of a health measurement scale. Journal of Clinical Epidemiology 1999;52:105-111.

371. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indices and rating scales. Journal of Clinical Epidemiology 1992;45:1201-1218.

372. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician 1983;32:307-317.

Table 1.1

Epidemiology of some neurological diseases [1]

Disease	For a health district of N = 250, 000	
	Incidence	Prevalence
Stroke	550	1500
Head injury	500	Unknown
Epilepsy	175	3900
Parkinson's disease	45	400
Multiple sclerosis	10	250
Motor neurone disease	5	15

[1] From Wade and Langton Hewer 1987 (3)

Table 1.2

Classification of health outcomes in neurology [1]

1	Physician-oriented outcomes	
	1.1	pathophysiological parameters of disease
	1.2	clinical end-points
2	Patient-oriented outcomes	
	2.1	aspects of health status
	2.2	health-related quality of life

[1] Adapted from Gill 1995 (34)

Table 1.3

Functional Independence Measure (FIM)

Scale	Subscale	Item
Motor	Self-care	Feeding
		Grooming
		Bathing
		Dressing upper body
		Dressing lower body
		Toileting
	Sphincter care	Bladder management
		Bowel management
	Transfer	Bed / chair transfer
		Toilet transfer
		Shower / tub transfer
	Locomotion	Walk
		stairs
Cognitive	Communication	Comprehension
		Expression
	Social cognition	Social interaction
		Problem solving
		Memory

Table 1.4

Functional Independence Measure + Functional Assessment Measure (FIM+FAM)

Scale	Subscale	Item
Motor	Self-care	Feeding
		Grooming
		Bathing
		Dressing upper body
		Dressing lower body
		Toileting
		Swallowing
	Sphincter care	Bladder management
		Bowel management
	Transfer	Bed / chair transfer
		Toilet transfer
		Shower / tub transfer
		Car transfer
	Locomotion	Walk
		stairs
		Community mobility
Cognitive	Communication	Comprehension
		Expression
		Reading
		Writing
		Speech intelligibility
	Psychosocial adjustment	Social interaction
		Emotional status
		Adjustment to limitations
		Employability
	Cognitive functions	Problem solving
		Memory
		Orientation
		Attention
		Safety judgement

Table 1.5

Response options for the FIM and FIM+FAM

Independent: **Another person is not required for the activity**

7 **Complete independence** - all of the tasks described as making up the activity are typically completed safely, without modification, assistive devices, or aids, and within a reasonable time.

6 **Modified independence** - one or more of the following may be true: requires an assistive device, activity takes more than a reasonable time, there are safety (risk) considerations.

Dependent: **Subject requires help from another person for either supervision or physical assistance in order for the activity to be performed, or it is not performed.**

Modified dependence: subject expends more than half (50%) of the effort required to complete the task.

5 **Supervision or setup** - subject requires no more help than standby, cueing or coaxing, without physical contact, or, helper sets up needed items or applies orthoses.

4 **Minimal contact assistance** - subject requires no more help than touching, and expends 75% or more of the effort.

3 **Moderate assistance** - subject requires more help than touching, or expends half (50%) or more (up to 75%) of the effort.

Complete dependence: subject expends less than half (50%) of the effort or the activity is not performed.

2 **Maximal assistance** - subject expends less than 50% of the effort, but at least 25%.

1 **Total assistance** - subject expends less than 25% of the effort.

Table 2.1

Health outcomes assessment

Measure	Method of administration	Site of administration	Assessment point
FIM	MDT [1]	All	A + D [2]
FIM+FAM	MDT	All	A + D
Barthel Index	MDT	NRU [3], RRU [4]	A + D
Modified Barthel Index	MDT	RNRU [5]	A + D
EDSS [6]	Neurologist [7]	NRU	A + D
OPCS [8]	Coordinator [9]	NRU, RRU	A + D
EDSS [10]	Neurologist [11]	NRU	A + D
LHS [12]	Self-report	NRU, RRU	A + D
SF-36 [13]	Self-report	NRU, RRU	A + D [14]
GHQ [15]	Self-report	NRU, RRU	A + D
MMSE [16]	Coordinator	NRU, RRU	A + D
Neuropsychological testing	Neuro-psychologist	NRU	A
Transition question	MDT	All	D

[1] Instrument rated by consensus opinion of treating multidisciplinary team.
[2] A = admission; D = discharge.
[3] Neurorehabilitation Unit, National Hospital for Neurology and Neurosurgery, London.
[4] Rehabilitation Research Unit, University of Southampton.
[5] Regional Neurorehabilitation Unit, Homerton Hospital, London.
[6] Expanded Disability Status Scale
[7] Rated by neurologist.
[8] Office of Population Censuses and Surveys Disability Scales
[9] Instrument rated at each site by the study coordinator.
[10] Expanded Disability Status Scale
[11] Rated by neurologist.
[12] London Handicap Scale
[13] Medical Outcomes Study 36-Item Short-Form Health Survey
[14] Administered by postal survey 4 weeks after discharge.
[15] General Health Questionnaire
[16] Mini-Mental State Examination

Table 3.1

Expected correlations between FIM and FIM+FAM scales and other measures

Measure	Scoring direction[2]	FIM / FIM+FAM scales[1] Total	Motor	Cog.	Rank order[3]
Barthel Index	+	+++	+++	++	M > T > C
Modified Barthel Index	+	+++	+++	++	M > T > C
EDSS[4]	-	- - -	- - -	- -	M > T > C
OPCS [5]	-	- - -	- - -	- -	M > T > C
LHS [6]	+	++	++	+	M > T > C
SF-36 PCS [7]	+	++	++	+	M > T > C
SF-36 MCS [8]	+	+	+	+	C > T > M
GHQ [9]	-	-	-	-	T = M = C
Mini Mental State Examination	+	++	+	+++	C > T > M
WAIS-R Verbal IQ	+	+	+	++	C > T > M
Halstead Book Category Test	-	-	-	- -	C > T > M
Wisconsin Card Sorting Test	+	+	+	++	C > T > M
Verbal and Spatial Reasoning Test	+	+	+	++	C > T > M
California Verbal Learning Test	+	+	+	++	C > T > M
Recognition Memory Test (verbal)	+	+	+	++	C > T > M

[1] Direction and number of + and - signs reflect the direction and magnitude of correlations; + / - = weak positive / negative correlation ($r < .30$); ++ / - - = moderate positive / negative correlation ($.30 < r < .70$); +++ / - - - = strong positive / negative correlation ($r > .70$).
[2] High scores indicate favourable (+) or unfavourable (-) health status.
[3] Indicates the rank order of correlations between FIM or FIM+FAM total (T), motor (M), and cognitive scales and the validating measure.
[4] Expanded Disability Status Scale
[5] Office of Population Censuses and Surveys Disability Scales
[6] London Handicap Scale
[7] Medical Outcomes Study 36-Item Short-Form Health Survey Physical Component Summary Score
[8] Medical Outcomes Study 36-Item Short-Form Health Survey Mental Component Summary Score
[9] General Health Questionnaire

Table 3.2

Characteristics of patient samples compared to 1994 local populations

		Clinical site								
	Total	NRU			RNRU			RRU		
Variable	Sample (*N* = 209)	Sample (*n* = 118)	Population (*n* = 138)	*p*[1]	Sample (*n* = 60)	Population (*n* = 62)	*p*	Sample (*n* = 31)	Population (*n* = 52)	*p*
Gender *n* (%)										
Male	106 (50.7)	55 (46.6)	66 (47.8)	NS	39 (65.0)	39 (62.9)	NS	13 (41.9)	37 (71.2)	< .01
Female	103 (49.3)	63 (53.4)	72 (52.2)		21 (35.0)	23 (37.1)		18 (58.1)	15 (28.8)	
Age (years)										
Mean (SD)	44.0 (14.4)	46 (14.8)	44 (15.3)	NS	39 (12.2)	37 (11.3)	NS	47 (14.7)	47 (16.2)	NS
Range	16 - 77	17 - 77	16 - 87		16 - 61	16 - 57		16 - 65	16 - 83	
Diagnosis *n* (%)										
Stroke	71 (34.0)	26 (22.0)	28 (20.3)	NS	26 (43.3)	28 (45.2)	NS	19 (61.3)	12 (23.1)	< .001
Multiple Sclerosis	64 (30.6)	64 (54.2)	59 (42.8)	NS	0 (0)	0 (0)	NS	0 (0)	5 (9.6)	NS
Head injury	33 (15.8)	3 (2.5)	0 (0)	< .05	24 (40.0)	24 (38.7)	NS	6 (19.4)	19 (36.5)	NS
Other	41 (19.6)	25 (21.2)	51 (37.0)	< .01	10 (16.7)	10 (16.1)	NS	6 (19.4)	16 (30.8)	NS
Length of stay (days)										
Mean (SD)	64.6 (68.8)	31 (24.5)	33 (27.2)	NS	128 (86.1)	128 (89.1)	NS	78 (52.2)	55 (59.7)	NS
Range	11 - 396	11 - 137	7 - 193		30 - 396	14 - 518		15 - 244	1 - 271	

[1] Tests of significance are based on Chi-squared tests for categorical variables and one-way ANOVA for continuous variables.

Table 3.3

Characteristics of stroke and MS patients

Variable	Diagnosis	
	Stroke (n = 71)	MS (n = 64)
Gender n (%)		
Male	40 (56.3)	23 (35.9)
Female	31 (43.7)	41 (64.1)
Age (years)		
Mean (SD)	49.8 (13.7)	43.0 (11.9)
Range	18 - 77	21 - 69
Length of stay (days)		
Mean (SD)	77.7 (55.3)	20.7 (11.8)
Range	11 - 312	11 - 100
Clinical site n (%)		
NRU	26 (36.6)	64 (100)
RNRU	26 (36.6)	0 (0)
RRU	19 (26.8)	0 (0)

Table 3.4

Comparison of FIM and FIM+FAM admission scores for patients in the reproducibility subsamples

Scale	Total Sample (N = 209) mean (SD)	Intra-rater reproducibility (n = 77) mean (SD)	Inter-rater reproducibility (n = 89) mean (SD)
FIM total	80.4 (29.3)	76.4 (30.4)	80.2 (31.0)
FIM motor	55.8 (22.6)	52.6 (23.6)	55.2 (23.6)
FIM cognitive	24.6 (9.3)	23.8 (9.8)	25.0 (9.4)
FIM+FAM total	135.8 (45.6)	129.8 (47.4)	136.0 (48.7)
FIM+FAM motor	68.4 (26.6)	64.5 (27.9)	68.0 (27.8)
FIM+FAM cognitive	67.4 (23.0)	65.3 (24.1)	68.0 (23.8)

Table 3.5

Descriptive statistics for FIM item scores at admission (*N* = 209)

Item	Response categories and endorsement frequencies (%)							Mean (SD)	% Floor effect	% Ceiling effect
	1	2	3	4	5	6	7			
Feeding	12.0	2.9	0.0	5.7	44.5	7.7	27.3	5.0 (1.8)	12.0	27.3
Grooming	12.0	4.8	6.2	12.9	16.3	11.0	36.8	5.0 (2.1)	12.0	36.8
Bathing	15.3	9.1	13.4	14.4	11.0	12.0	24.9	4.3 (2.2)	15.3	24.9
Dressing upper body	12.0	7.2	8.6	15.8	8.1	21.5	26.8	4.7 (2.1)	12.0	26.8
Dressing lower body	28.2	9.1	9.6	13.9	8.6	18.2	12.4	3.7 (2.2)	28.2	12.4
Toileting	22.5	8.1	10.0	8.1	1.9	23.4	25.8	4.3 (2.4)	22.5	25.8
Bladder management	19.6	4.8	5.3	5.3	7.7	17.7	39.7	4.9 (2.4)	19.6	39.7
Bowel management	13.4	1.4	3.3	5.7	3.8	25.8	46.4	5.5 (2.1)	13.4	46.4
Bed transfer	17.2	7.2	13.9	10.0	7.2	25.4	19.1	4.4 (2.2)	17.2	19.1
Toilet transfer	17.7	6.7	12.0	10.5	4.8	38.3	10.0	4.3 (2.1)	17.7	10.0
Shower/tub transfer	29.2	9.1	12.4	21.5	9.1	16.7	1.9	3.3 (1.9)	29.2	1.9
Walk	34.9	5.7	4.3	7.2	16.3	23.9	7.7	3.7 (2.3)	34.9	7.7
Stairs	56.0	4.3	3.8	6.2	11.0	16.3	2.4	2.7 (2.1)	56.0	2.4
Comprehension	7.2	4.8	5.3	7.2	13.4	14.8	47.4	5.5 (1.9)	7.2	47.4
Expression	11.5	5.3	5.3	8.1	15.3	13.4	41.1	5.2 (2.1)	11.5	41.1
Social interaction	10.0	7.2	10.0	7.2	8.6	16.7	40.2	5.1 (2.1)	10.0	40.2
Problem solving	17.7	12.9	12.9	10.0	13.4	13.9	19.1	4.1 (2.2)	17.7	19.1
Memory	12.4	12.9	7.2	7.7	7.2	10.0	42.6	4.8 (2.3)	12.4	42.6

Table 3.6

Descriptive statistics for FIM scale scores at admission (*N* = 209)

FIM scale	No. items	Range of scores: Scale	Range of scores: Sample	Scale mid-point	Sample mean (SD)	Floor effect %	Ceiling effect %	Skew-ness
Total	18	18 - 126	18 - 122	72	80.4 (29.3)	2.4	0	- .573
Motor	13	13 - 91	13 - 91	52	55.8 (22.6)	4.8	0.5	- .417
Cognitive	5	5 - 35	5 - 35	20	24.6 (9.3)	4.3	12.4	- .706

Table 3.7
Reliability estimates for FIM scales (*N* = 209)

FIM scale	Internal consistency			Reproducibility		95% confidence interval for individual scores [1]	
	Item-total correlation range (mean)	Alpha (LL95%CI) [2]	Homogeneity coefficient	Intra-rater ICC [3] (LL95%CI) *n* = 77	Inter-rater ICC (LL95%CI) *n* = 89	Cross-sectional [4]	Longitudinal [5]
Total	.60 - .87 (.73)	.96 (.95)	.56	.98 (.96)	.98 (.97)	11.5	8.1
Motor	.63 - .92 (.79)	.96 (.95)	.64	.98 (.97)	.98 (.97)	8.9	6.3
Cognitive	.77 - .85 (.80)	.92 (.92)	.71	.95 (.92)	.94 (.92)	5.2	4.1

[1] +/- 1.96 SEM
[2] Lower limit of 95% confidence interval
[3] Intraclass Correlation Coefficient
[4] Uses alpha coefficient as reliability estimate in calculation of standard error of measurement.
[5] Uses intra-rater reproducibility coefficient as reliability estimate in calculation of standard error of measurement.

Table 3.8

Intercorrelations between FIM scales (N = 209)

	FIM scale	
FIM scale	Total	Motor
Motor	.97	
Cognitive	.79	.61

Table 3.9

Correlations between the FIM, other outcome measures, and age

	Health outcome measures [1]								
	Disability				Handicap	Health status		Psychological distress	
FIM scale	BI[2]	MBI[3]	EDSS[4]	OPCS[5]	LHS[6]	SF-36 PCS[7]	SF-36 MCS[8]	GHQ[9]	Age [10]
Total	.95	.93	- .84	- .82	.32	.26	.10	- .13	- .06
Motor	.97	.96	- .86	- .84	.35	.30	.10	- .15	- .09
Cognitive	.57	.65	- .45	- .44	.11	.04	.08	- .01	- .03

[1] Sample sizes differ as not all instruments were administered at each study site. Also, the EDSS was only administered to MS patients.
[2] Barthel Index (n = 149)
[3] Modified Barthel Index (n = 60)
[4] Kurtzke Extended Disability Status Scale (n = 64)
[5] Office of Population Censuses and Surveys Overall Weighted Disability Severity Score (n =69)
[6] London Handicap Scale (n = 121)
[7] SF-36 Physical Component Summary Score (n = 123)
[8] SF-36 Mental Component Summary Score (n = 123)
[9] General Health Questionnaire (n = 85)
[10] N = 209

Table 3.10

Correlations between the FIM and neuropsychological measures

	Neuropsychological measures [1]						
	Global decline		Reasoning			Memory	
FIM scale	MMSE [2]	WAIS-VIQ [3]	HBCT [4]	WCST [5]	VESPAR [6]	CVLT [7]	RMT [8]
Total	.49	.35	-.34	.52	.53	.55	.59
Motor	.32	.27	-.27	.41	.40	.48	.56
Cognitive	.76	.51	-.42	.68	.75	.61	.50

[1] Sample sizes differ as not all instruments were administered at each study site. Also, the VESPAR was only administered to MS patients.
[2] Mini-Mental State Examination (n = 90)
[3] Wechsler Adult Intelligence Test (revised version) - Verbal IQ (n = 60)
[4] Halstead Booklet Category Test (n = 44)
[5] Wisconsin Card Sorting Test (n = 40)
[6] Verbal and Spatial Reasoning Test Total Score (n = 37)
[7] California Verbal Learning Test (n = 52)
[8] Recognition Memory Test - Visual Version (n = 49)

Table 3.11

Mean FIM change scores for different levels of staff-rated improvement in disability

FIM scale	Staff-rated improvement in disability: None *n* = 17	Minimal *n* = 42	Moderate *n* = 75	Marked *n* = 45	*F*	*p*
Total	- 1.6	- 6.2	- 14.2	- 31.6	31.4	< .001
Motor	- 2.2	- 5.9	- 11.9	- 27.2	30.2	< .001
Cognitive	.6	- .3	- 2.3	- 4.4	8.7	< .001

Table 3.12

Mean FIM change scores and standard deviations for stroke and MS patients

FIM scale	Mean change score (SD)		*p*
	Stroke (*n* = 62)	MS (*n* = 64)	
Total	- 19.3 (17.5)	- 6.4 (7.9)	< .001
Motor	- 16.8 (15.2)	- 6.5 (6.7)	< .001
Cognitive	- 2.4 (4.7)	.2 (3.2)	.130

Table 3.13

Responsiveness of the FIM (n = 194)

	Mean score (SD)			Responsiveness
FIM scale	Admission	Discharge	Change[1]	Effect size
Total	82.2 (28.6)	97.0 (27.8)	- 14.8 (17.0)	- .52
Motor	57.0 (22.1)	70.0 (21.9)	- 12.8 (14.5)	- .58
Cognitive	25.2 (9.1)	27.2 (8.0)	- 2.0 (4.6)	- .22

[1] All change scores are statistically significant ($p < .001$).

Table 3.14

Relative responsiveness of disability measures

Measure	n^1	Effect size
FIM total	194	- .52
FIM motor	194	- .58
FIM cognitive	194	- .22
FIM+FAM total	194	- .45
FIM+FAM motor	194	- .57
FIM+FAM cognitive	194	- .24
Barthel Index	136	- .56
Modified Barthel Index	57	- .67
EDSS [2]	64	.06
OPCS [3]	60	.38

[1] Sample sizes differ as not all instruments were administered at each site. Also, the EDSS was only administered to MS patients.
[2] Expanded Disability Status Scale
[3] Office of Population Censuses and Surveys Overall Weighted Disability Severity Score

Table 3.15

Descriptive statistics for FIM+FAM item scores at admission (*N* = 209)

Item	Response categories and endorsement frequencies (%)							Mean (SD)	% Floor effect	% Ceiling effect
	1	2	3	4	5	6	7			
Feeding	12.0	2.9	0.0	5.7	44.5	7.7	27.3	5.0 (1.8)	12.0	27.3
Grooming	12.0	4.8	6.2	12.9	16.3	11.0	36.8	5.0 (2.1)	12.0	36.8
Bathing	15.3	9.1	13.4	14.4	11.0	12.0	24.9	4.3 (2.2)	15.3	24.9
Dressing upper body	12.0	7.2	8.6	15.8	8.1	21.5	26.8	4.7 (2.1)	12.0	26.8
Dressing lower body	28.2	9.1	9.6	13.9	8.6	18.2	12.4	3.7 (2.2)	28.2	12.4
Toileting	22.5	8.1	10.0	8.1	1.9	23.4	25.8	4.3 (2.4)	22.5	25.8
Swallowing	3.3	2.9	1.0	2.4	6.7	12.9	70.8	6.3 (1.5)	3.3	70.8
Bladder care	19.6	4.8	5.3	5.3	7.7	17.7	39.7	4.9 (2.4)	19.6	39.7
Bowel care	13.4	1.4	3.3	5.7	3.8	25.8	46.4	5.5 (2.1)	13.4	46.4
Bed transfer	17.2	7.2	13.9	10.0	7.2	25.4	19.1	4.4 (2.2)	17.2	19.1
Toilet transfer	17.7	6.7	12.0	10.5	4.8	38.3	10.0	4.3 (2.1)	17.7	10.0
Shower/tub transfer	29.2	9.1	12.4	21.5	9.1	16.7	1.9	3.3 (1.9)	29.2	1.9
Car transfer	26.8	10.5	18.7	14.4	10.0	10.0	9.6	3.4 (2.0)	26.8	9.6
Walk	34.9	5.7	4.3	7.2	16.3	23.9	7.7	3.7 (2.3)	34.9	7.7
Stairs	56.0	4.3	3.8	6.2	11.0	16.3	2.4	2.7 (2.1)	56.0	2.4
Community mobility	38.8	16.7	6.2	14.4	7.2	8.6	8.1	2.9 (2.1)	38.8	8.1
Comprehension	7.2	4.8	5.3	7.2	13.4	14.8	47.4	5.5 (1.9)	7.2	47.4
Expression	11.5	5.3	5.3	8.1	15.3	13.4	41.1	5.2 (2.1)	11.5	41.1
Reading	11.0	0.5	3.8	9.6	14.8	20.1	40.2	5.4 (1.9)	11.0	40.2
Writing	16.3	5.3	11.0	13.4	8.6	12.9	32.5	4.6 (2.2)	16.3	32.5
Speech intelligibility	4.8	3.8	11.5	2.4	9.6	20.6	47.4	5.6 (1.8)	4.8	47.4
Social interaction	10.0	7.2	10.0	7.2	8.6	16.7	40.2	5.1 (2.1)	10.0	40.2
Emotional status	8.1	7.7	13.9	14.8	13.4	24.4	17.7	4.6 (1.9)	8.1	17.7
Adjustment to limitation	12.0	17.7	8.6	8.1	18.7	14.4	16.7	4.2 (2.0)	12.0	16.7
Employability	35.4	16.7	9.1	11.0	11.0	14.4	2.4	3.0 (2.0)	35.4	2.4
Problem solving	17.7	12.9	12.9	10.0	13.4	13.9	19.1	4.1 (2.2)	17.7	19.1
Memory	12.4	12.9	7.2	7.7	7.2	10.0	42.6	4.8 (2.3)	12.4	42.6
Orientation	8.1	4.3	5.3	4.3	5.3	7.2	65.6	5.8 (2.0)	8.1	65.6
Attention	9.1	8.6	12.9	7.2	9.6	7.7	45.0	5.0 (2.2)	9.1	45.0
Safety judgement	9.6	11.0	9.6	10.0	22.5	13.4	23.9	4.6 (2.0)	9.6	23.9

Table 3.16

Descriptive statistics for FIM+FAM scale scores at admission (*N* = 209)

FIM+FAM scale	No. items	Range of scores: Scale	Range of scores: Sample	Scale mid-point	Sample mean (SD)	Floor effect %	Ceiling effect %	Skew-ness
Total	30	30 - 210	30 - 204	120	135.8 (45.6)	.5	0	- .610
Motor	16	16 - 112	16 - 110	64	68.4 (26.6)	1.4	0	- .376
Cognitive	14	14 - 98	14 - 98	56	67.4 (23.0)	1.4	1.9	- .702

Table 3.17
Reliability estimates for the FIM+FAM (*N* = 209)

	Internal consistency			Reproducibility		95% confidence intervals for individual scores [1]	
FIM+FAM scale	Item-total correlation range (mean)	Alpha (LL95%CI) [2]	Homogeneity coefficient	Intra-rater ICC [3] (LL95%CI) *n* = 77	Inter-rater ICC (LL95%CI) *n* = 89	Cross-sectional [4]	Longitudinal [5]
Total	.55 - .85 (.71)	.97 (.96)	.53	.98 (.97)	.98 (.97)	15.5	12.6
Motor	.57 - .91 (.77)	.96 (.95)	.61	.98 (.97)	.98 (.97)	10.4	7.4
Cognitive	.63 - .86 (.77)	.96 (.95)	.62	.97 (.95)	.96 (.93)	9.0	7.8

[1] +/- 1.96 SEM
[2] Lower limit 95% confidence interval
[3] Intraclass Correlation Coefficient
[4] Uses alpha coefficient as reliability estimate in the calculation of standard error of measurement.
[5] Uses intra-rater reproducibility coefficient as reliability estimate in the calculation of standard error of measurement.

Table 3.18

Intercorrelations between FIM+FAM scales (*N* = 209)

	FIM+FAM scale	
FIM+FAM scale	Total	Motor
Motor	.93	
Cognitive	.91	.69

Table 3.19

Correlations between the FIM+FAM, other outcome measures, and age

	Health outcome measures [1]								
	Disability				Handicap	Health status		Psychological distress	
FIM+FAM scale	BI[2]	MBI[3]	EDSS[4]	OPCS[5]	LHS[6]	SF-36 PCS[7]	SF-36 MCS[8]	GHQ[9]	Age [10]
Total	.90	.89	- .79	- .77	.32	.24	.12	- .13	- .04
Motor	.97	.95	- .86	- .84	.36	.29	.10	- .14	- .09
Cognitive	.63	.70	- .48	- .50	.19	.10	.13	- .07	.03

[1] Sample sizes differ as instruments were administered at different combinations of study sites and the EDSS is MS specific.
[2] Barthel Index (*n* = 149)
[3] Modified Barthel Index (*n* = 60)
[4] Kurtzke Extended Disability Status Scale (*n* = 64)
[5] Office of Population Censuses and Surveys Overall Weighted Disability Severity Score (*n* =69)
[6] London Handicap Scale (*n* = 121)
[7] SF-36 Physical Component Summary Score (*n* = 123)
[8] SF-36 Mental Component Summary Score (*n* = 123)
[9] General Health Questionnaire (*n* = 85)
[10] *N* = 209

Table 3.20

Correlations between the FIM+FAM and neuropsychological measures

	Neuropsychological measures [1]						
	Global decline		Reasoning			Memory	
FIM+FAM scale	MMSE [2]	WAIS-VIQ [3]	HBCT [4]	WCST [5]	VESPAR [6]	CVLT [7]	RMT [1]
Total	.58	.42	- .40	.59	.60	.58	.58
Motor	.35	.29	- .28	.45	.43	.50	.56
Cognitive	.75	.54	- .49	.69	.76	.61	.50

[1] Sample sizes differ as different instruments were administered at different combinations of study sites and the VESPAR is MS specific.
[2] Mini-Mental State Examination (n = 90)
[3] Wechsler Adult Intelligence Test - Verbal IQ (n = 60)
[4] Halstead Booklet Category Test (n = 44)
[5] Wisconsin Card Sorting Test (n = 40)
[6] Verbal and Spatial Reasoning Test Total Score (n = 37)
[7] California Verbal Learning Test (n = 52)

Table 3.21

Mean FIM+FAM change scores for different levels of staff-rated improvement in disability

FIM+FAM scale	Staff-rated improvement in disability: None *n* = 17	Minimal *n* = 42	Moderate *n* = 75	Marked *n* = 45	*F*	*p*
Total	.1	- 8.5	- 20.2	- 43.4	29.9	.001
Motor	- 3.2	- 6.6	- 14.2	- 31.3	29.9	.001
Cognitive	3.3	- 1.9	- 5.9	- 12.1	13.3	.001

Table 3.22

Mean FIM+FAM change scores and standard deviations for stroke and MS patients

	Mean change score (SD)		
FIM+FAM scale	Stroke (*n* = 62)	MS (*n* = 64)	*p*
Total	- 25.3 (25.8)	- 8.7 (11.3)	.001
Motor	- 19.6 (17.6)	- 7.3 (8.0)	.001
Cognitive	- 5.7 (11.3)	- 1.4 (6.4)	.097

Table 3.23

Responsiveness of the FIM+FAM (n = 194)

	Mean score (SD)			Responsiveness
FIM+FAM scale	Admission	Discharge	Change[2]	Effect size
Total	138.7 (44.5)	158.8 (42.8)	- 20.1 (24.3)	- .45
Motor	69.8 (25.9)	84.7 (25.5)	- 14.9 (16.7)	- .57
Cognitive	68.9 (22.5)	74.2 (20.5)	- 5.3 (10.7)	- .24

[1] Recognition Memory Test - Visual Version (n = 49)
[2] All change scores are statistically significant (p < .001).

Table 3.24

The FIM in stroke and MS patients: descriptive statistics, reliability estimates, and intercorrelations between scales

Descriptive statistics	FIM total		FIM motor		FIM cognitive	
	Stroke [1]	MS [2]	Stroke	MS	Stroke	MS
Possible range (scale mid-point)	18 - 126 (72)	18 - 126 (72)	13 - 91 (52)	13 - 91 (52)	5 - 35 (20)	5 - 35 (20)
Actual range	21 - 122	24 - 121	13 - 87	13 - 88	6 - 35	11 - 35
Mean score (SD)	76.4 (26.5)	91.2 (24.1)	53.3 (20.6)	61.7 (20.4)	23.1 (8.9)	29.6 (6.0)
Floor / ceiling effect %	0 / 0	0 / 0	2.8 / 0	1.6 / 0	0 / 7.0	0 / 15.6
Reliability estimates						
Homogeneity coefficient	.52	.47	.62	.57	.67	.54
Item-total correlation - range	.54 - .86	.34 - .86	.55 - .92	.32 - .89	.72 - .82	.60 - .75
Alpha	.95	.94	.95	.95	.91	.83
Intra-rater reproducibility[3] (ICC[4])	.98	.92	.99	.93	.94	.92
Inter-rater reproducibility[5] (ICC)	.97	.96	.97	.98	.95	.88
Intercorrelations between scales						
FIM total	1.00	1.00	.96	.98	.76	.70
FIM motor	.96	.98	1.00	1.00	.55	.54
FIM cognitive	.76	.70	.55	.54	1.00	1.00

[1] n = 71 unless specified.
[2] n = 64 unless specified.
[3] n = 29 for stroke, n = 17 for MS.
[4] Intraclass Correlation Coefficient
[5] n = 29 for stroke, n = 34 for MS.

Table 3.25

The FIM in stroke and MS patients: external construct validity and responsiveness

	n		FIM total		FIM motor		FIM cognitive	
External construct validity	Stroke	MS	Stroke	MS	Stroke	MS	Stroke	MS
Barthel Index	45	64	.95	.94	.97	.96	.57	.50
LHS [1]	33	58	.15	.42	.22	.43	- .10	.22
SF-36 PCS [2]	34	58	.13	.14	.16	.14	- .01	.09
SF-36 MCS [3]	34	58	.19	.28	.24	.30	- .03	.11
Age	71	64	- .13	- .05	- .15	- .04	- .02	- .09
Responsiveness								
Mean change score (SD)	62	64	-19.3 (17.5)	-6.4 (7.9)	-16.8 (15.2)	-6.5 (6.7)	-2.4 (4.7)	.2 (3.2)
Effect size	62	64	- .75	- .27	- .84	- .32	- .29	.03

[1] London Handicap Scale
[2] SF-36 Physical Component Summary Score
[3] SF-36 Mental Component Summary Score

Table 3.26

The FIM+FAM in stroke and MS patients: descriptive statistics, reliability estimates, and intercorrelations between scales

	FIM+FAM total		FIM+FAM motor		FIM+FAM cognitive	
Descriptive statistics	Stroke [1]	MS [2]	Stroke	MS	Stroke	MS
Possible range (scale mid-point)	30 - 210 (120)	30 - 210 (120)	16 - 112 (64)	16 - 112 (64)	14 - 98 (56)	14 - 98 (56)
Actual range	35 - 203	54 - 203	16 - 108	20 - 106	19 - 96	34 - 98
Mean score (SD)	129.5 (41.2)	154.6 (35.2)	65.1 (24.2)	76.1 (23.6)	64.4 (21.0)	78.5 (15.2)
Floor / ceiling effect %	0 / 0	0 / 0	1.4 / 0	0 / 0	0 / 0	0 / 1.6
Reliability estimates						
Homogeneity coefficient	.49	.41	.59	.54	.56	.44
Item-total correlation - range	.47 - .81	.27 - .82	.47 - .92	.34 - .89	.55 - .84	.36 - .87
Alpha	.97	.95	.96	.95	.95	.91
Intra-rater reproducibility[3] (ICC[4])	.98	.92	.99	.93	.94	.92
Inter-rater reproducibility[5] (ICC)	.97	.96	.97	.98	.95	.88
Intercorrelations between scales						
FIM+FAM total	1.00	1.00	.92	.94	.90	.85
FIM+FAM motor	.92	.94	1.00	1.00	.66	.62
FIM+FAM cognitive	.90	.85	.66	.62	1.00	1.00

[1] n = 71 for stroke unless specified.
[2] n = 64 for MS unless specified.
[3] n = 29 for stroke, n = 17 for MS.
[4] Intraclass Correlation Coefficient
[5] n = 29 for stroke, n = 34 for MS.

Table 3.27

The FIM+FAM in stroke and MS patients: external construct validity and responsiveness

External construct validity	*n*		FIM+FAM total		FIM+FAM motor		FIM+FAM cognitive	
	Stroke	MS	Stroke	MS	Stroke	MS	Stroke	MS
Barthel Index	45	64	.92	.88	.97	.96	.67	.54
LHS [1]	33	58	.15	.41	.24	.44	- 01	.26
SF-36 PCS [2]	34	58	.14	.14	.16	.13	.08	.09
SF-36 MCS [3]	34	58	.17	.27	.25	.29	.01	.17
Age	71	64	- .12	- .07	- .15	- .03	- .06	- .01
Responsiveness								
Mean change score (SD)	62	64	-25.3 (25.8)	-8.7 (11.3)	-19.6 (17.6)	- 7.3 (8.0)	- 5.7 (11.3)	- 1.4 (6.4)
Effect size	62	64	- .66	- .28	- .83	- .31	- .29	- .09

[1] London Handicap Scale
[2] SF-36 Physical Component Summary Score
[3] SF-36 Mental Component Summary Score

Table 3.28

Descriptive statistics for the FIM, FIM+FAM, and Barthel Index (n = 149)

Variable	Global disability		Motor disability			Cognitive disability	
	FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
Number of items	18	30	13	16	10	5	14
Possible range	18 - 126	30 - 210	13 - 91	16 -112	0 - 20	5 - 35	14 - 98
Actual range	19 - 122	32 - 204	13 - 91	17 - 110	0 - 20	6 - 35	15 - 98
Scale mid-point	72	120	52	64	10	20	56
Sample mean (SD)	85.0 (26.7)	144.8 (40.5)	57.7 (21.4)	71.0 (25.2)	11.5 (5.9)	27.2 (7.8)	73.8 (19.0)
Floor effect %	0	0	2.7	0	2.7	0	0
Ceiling effect %	0	0	.7	0	5.4	16.1	2.7
Skewness	- .577	- .603	- .457	- .396	- .257	- 1.040	- .896

Table 3.29

Reliability estimates for the FIM, FIM+FAM, and Barthel Index

Variable	Global disability		Motor disability			Cognitive disability	
	FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
Internal consistency (n = 149)							
Homogeneity coefficient	.51	.46	.60	.57	.51	.63	.51
Item-total correlation[1] - range	.53 - .87	.40 - .82	.56 - .91	.51 - .90	.46 - .84	.69 - .80	.49 - .87
Cronbach's alpha coefficient	.95	.96	.95	.96	.94	.89	.91
Reproducibility (ICC[2])							
Intra-rater (n = 77)	.98	.98	.98	.98	N / A	.95	.97
Inter-rater (n = 89)	.98	.98	.98	.98	N / A	.94	.96

[1] Corrected for overlap.
[2] Intraclass Correlation Coefficient

Table 3.30

Intercorrelations between FIM and FIM+FAM scales and the Barthel Index (*n* = 149)

Scale	Global disability		Motor disability			Cognitive disability	
	FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
FIM total	1.00	**.98** [1]	.97	.97	.95	.75	.80
FIM motor	.97	.92	1.00	**.996**	**.97**	.58	.64
FIM cognitive	.75	.83	.58	.60	.57	1.00	**.97**
FIM+FAM total	**.98**	1.00	.92	.94	.90	.83	.89
FIM+FAM motor	.97	.94	**.996**	1.00	.97	.60	.67
FIM+FAM cognitive	.80	.89	.64	.67	.63	**.97**	1.00
Barthel Index	.95	.90	**.97**	**.97**	1.00	.57	.63

[1] Values in **bold** indicate correlations between scales which purport to measure the same aspect of disability.

Table 3.31

Correlations between the FIM, FIM+FAM, Barthel Index and other outcome measures and age

Other measures			Global disability		Motor disability			Cognitive disability	
Construct	Scale	*n*[1]	FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
Disability	EDSS [2]	64	- .84	- .79	- .87	- .86	- .89	- .45	- .48
	OPCS [3]	69	- .82	- .77	- .84	- .84	- .84	- .43	- .50
Handicap	LHS [4]	121	.32	.32	.35	.36	.33	.11	.19
Health status	SF-36 PCS [5]	122	.26	.24	.30	.29	.30	.04	.10
	SF-36 MCS [6]	122	.10	.12	.10	.10	.11	.08	.13
Psychological distress	GHQ [7]	85	- .13	- 13	- .15	- .14	- .14	.01	- .07
Other	Age	149	- .13	- .13	- .15	- .15	- .12	- .06	- .12

[1] Sample sizes differ as not all measures were administered at each clinical site.
[2] Expanded Disability Status Scale
[3] Office of Population Censuses and Surveys Overall Weighted Disability Severity Score
[4] London Handicap Scale
[5] SF-36 Physical Component Summary Score
[6] SF-36 Physical Component Summary Score
[7] General Health Questionnaire

Table 3.32

Correlations between the FIM, FIM+FAM, Barthel Index and neuropsychological measures

Neuropsychological measure			Global disability		Motor disability			Cognitive disability	
Construct	Scale	*n* [1]	FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
Global decline	MMSE [2]	90	.49	.58	.32	.35	.36	.76	.75
	WAIS-VIQ [3]	43	.35	.42	.27	.29	.28	.51	.54
Reasoning	HBCT [4]	44	- .34	- .40	- .27	- 28	- .27	- .42	- .49
	WCST [5]	40	.52	.59	.41	.45	.35	.68	.69
	VESPAR [6]	37	.53	.60	.40	.43	.38	.75	.76
Memory	CVLT [7]	52	.55	.58	.48	.50	.45	.61	.61
	RMT [1]	49	.53	.56	.43	..46	.36	.63	.59

[1] Sample sizes differ as not all measures were administered at each study site. Also, the VESPAR was only administered to MS patients.
[2] Mini-Mental State Examination
[3] Wechsler Adult Intelligence Test - Verbal IQ
[4] Halstead Booklet Category Test
[5] Wisconsin Card Sorting Test
[6] Verbal and Spatial Reasoning Test Total Score
[7] California Verbal Learning Test

Table 3.33

Relative precision estimates for the FIM, FIM+FAM, and Barthel Index

Staff-rated improvement in disability	*n*	Global disability		Motor disability			Cognitive disability	
		FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
Minimal	30	- 5.4	- 7.7	- 5.5	- 6.1	- 2.1	.10	- 1.6
Moderate	50	- 11.4	- 15.8	- 9.7	- 11.4	- 2.8	- 1.7	- 4.4
Marked	28	- 30.2	- 40.1	-26.9	- 30.9	- 7.0	- 3.4	- 9.3
F - statistic		30.12	24.50	33.19	32.42	20.40	4.84	5.25
p		.0000	.0000	.0000	.0000	.0000	.0098	.0067
Relative precision[1]		1.0	.81	1.0	.98	.61	1.0	1.08

[1] Defined, relative to competing FIM scale, in terms of the degree to which a scale is able to detect group differences. Calculated by dividing *F* - statistic of competing scales by *F* - statistic of appropriate FIM scale.

Table 3.34

Responsiveness of the FIM, FIM+FAM, and Barthel Index (*n* = 136)

Mean score (SD)	Global disability		Motor disability			Cognitive disability	
	FIM total	FIM+FAM total	FIM motor	FIM+FAM motor	Barthel Index	FIM cognitive	FIM+FAM cognitive
Admission	87.2 (25.4)	148.7 (38.0)	59.1 (20.6)	72.7 (24.2)	11.9 (5.7)	28.1 (7.0)	76.0 (17.1)
Discharge	99.4 (24.9)	164.7 (37.0)	70.1 (20.6)	85.4 (23.9)	15.1 (5.4)	29.3 (6.2)	79.3 (15.8)
Change[2]	- 12.2 (15.3)	- 15.9 (21.9)	- 11.1 (13.0)	- 12.6 (15.0)	- 3.2 (3.7)	- 1.2 (4.4)	- 3.3 (9.9)
Responsiveness Effect size	- .48	- .42	- .54	- .52	- .56	- .17	- .19

[1] Recognition Memory Test - Visual Version
[2] All change scores are statistically significant at $p = < .001$ level except for the FIM cognitive scale which is statistically significant at $p = .002$.

Table 4.1

Measurement model of the FIM

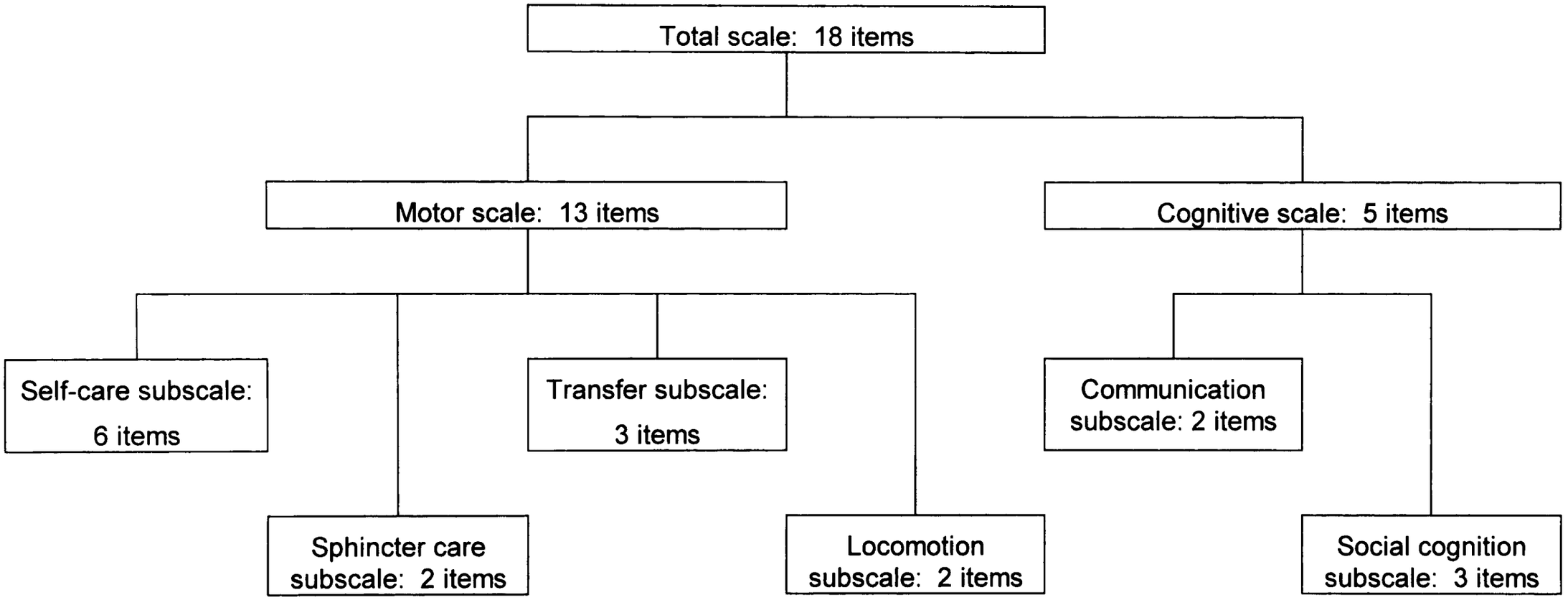


Table 4.2

Measurement model of the FIM+FAM

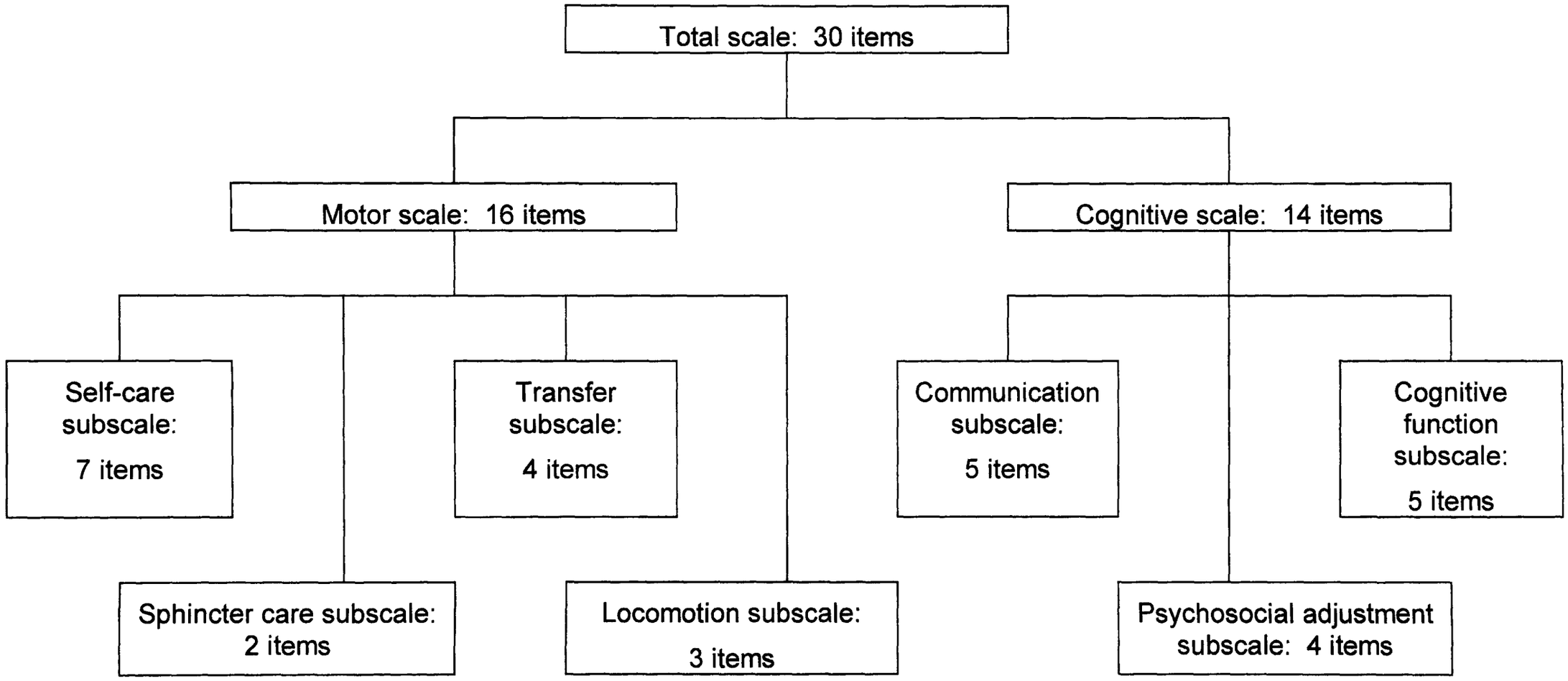


Table 4.3

The FIM: internal consistency and intercorrelations between scales and subscales (*N* = 209)

FIM scale	Internal consistency		Intercorrelations between FIM scales and subscales (alphas in parentheses)								
			FIM scale			FIM subscale					
	Item-total correlation	Item-inter correlation (mean)	T	M	C	Sc	Sp	Tr	L	Com	Soc
Total (T)	.60 - .87	.21 - .96 (.56)	(.96)								
Motor (M)	.63 - .92	.29 - .96 (.64)	.97	(.96)							
Cognitive (C)	.77 - .85	.60 - .87 (.71)	.79	.61	(.92)						
FIM subscale											
Self care (Sc)	.73 - .87	.58 - .86 (.73)	.96	.97*	.65	(.94)					
Sphincter (Sp)	.67	.67	.76	.77*	.51	.70	(.80)				
Transfer (Tr)	.82 - .94	.81 - .96 (.86)	.90	.95*	.51	.89	.68	(.95)			
Locomotion (L)	.68	.68	.69	.75*	.34	.64	.40	.71	(.81)		
Communication (Com)	.87	.87	.73	.57	.91*	.61	.44	.48	.37	(.93)	
Social cognition (Soc)	.76 - .84	.69 - .79 (.74)	.75	.58	.96*	.62	.51	.47	.29	.75	(.90)

* Subscale-own scale correlations.

Table 4.4

The FIM+FAM: internal consistency and intercorrelations between scales and subscales (N = 209)

FIM+FAM scale	Internal consistency		Intercorrelations between FIM+FAM scales and subscales (alphas in parentheses)									
			FIM+FAM scale			FIM+FAM subscale						
	Item-total correlation	Item-inter correlation (mean)	T	M	C	Sc	Sp	Tr	L	Com	Pa	Cf
Total (T)	.55 - .85	.15 - .96 (.53)	(.97)									
Motor (M)	.57 - .91	.29 - .96 (.61)	.93	(.96)								
Cognitive (C)	.63 - .86	.38 - .87 (.62)	.91	.69	(.96)							
FIM+FAM subscale												
Self care (Sc)	.61 - .86	.43 - .86 (.68)	.92	.96*	.71	(.94)						
Sphincter (Sp)	.67	.67	.72	.75*	.55	.71	(.80)					
Transfer (Tr)	.83 - .94	.76 - .96 (.83)	.84	.95*	.57	.88	.66	(.95)				
Locomotion (L)	.71 - .78	.64 - .74 (.69)	.64	.76*	.40	.63	.40	.74	(.87)			
Communication (Com)	.75 - .91	.63 - .87 (.74)	.82	.65	.89*	.68	.48	.55	.40	(.93)		
Psychosocial adj't (Pa)	.61 - .79	.41 - .71 (.61)	.83	.63	.92*	.64	.51	.52	.35	.71	(.86)	
Cognitive functions (Cf)	.81 - .85	.71 - .80 (.76)	.83	.62	.93*	.63	.53	.50	.33	.70	.84	(.94)

* Subscale-own scale correlations.

Table 4.5
Item-scale and item-subscale correlations for the FIM (*N* = 209)

		FIM scale		FIM subscale					
FIM subscale	FIM item	Motor	Cognitive	SC	Sp	Tr	Lo	Com	Soc
Self-care (SC)	Feeding	.70*[1]	.48	.73*	.52	***.63***[2]	.50	.48	.43
	Grooming	.80*	***.66***	.85*	.62	***.73***	.50	.59	.64
	Bathing	.86*	.65	.87*	.63	***.82***	.62	.60	.62
	Dressing upper body	.79*	.63	.82*	.56	***.74***	.56	.58	.60
	Dressing lower body	.88*	.51	.84*	.62	***.87***	.68	.48	.48
	Toileting	.89*	.53	.84*	***.71***	***.88***	.64	.49	.51
Sphincter-care	Bladder management	.65*	.46	***.63***	.67*	***.62***	.40	.39	.45
(Sp)	Bowel management	.65*	.49	***.64***	.67*	***.62***	.38	.42	.48
Transfer (Tr)	Bed transfer	.91*	.49	***.87***	.65	.93*	.72	.47	.45
	Toilet transfer	.92*	.50	***.88***	.68	.94*	.70	.46	.48
	Shower/tub transfer	.83*	.47	***.79***	.60	.82*	.66	.46	.44
Locomotion	Walk	.63*	.38	***.59***	.33	***.63***	.68*	.41	.32
(Lo)	Stairs	.70*	.30	***.63***	.46	***.71***	.68*	.32	.25
Communication	Comprehension	.51	.85*	.53	.41	.42	.35	.87*	***.77***
(Com)	Expression	.60	.77*	.63	.45	.51	.42	.87*	.70
Social cognition	Social interaction	.54	.80*	.57	.49	.44	.29	***.72***	.76*
(Soc)	Problem solving	.54	.81*	.58	.43	.45	.32	.67	.84*
	Memory	.50	.79*	.54	.47	.42	.25	***.67***	.79*

[1] Corrected item-own scale / subscale correlation.
[2] Values in **bold** indicate item-other subscale correlations which do not satisfy criteria for definite scaling success.

Table 4.6

Item-scale and item-subscale correlations for the FIM+FAM (*N* = 209)

FIM+FAM subscale	FIM+FAM item	FIM+FAM scale		FIM+FAM subscale						
		Mot	Cog	Sc	Sp	Tr	Lo	Co	Pa	Cf
Self-care (Sc)	Feeding	.71*[1]	.54	.77*	.52	***.63***[2]	.54	.57	.48	.43
	Grooming	.80*	***.71***	.85*	.62	***.73***	.58	.66	.64	.65
	Bathing	.87*	.69	.86*	.63	***.82***	.70	.64	.63	.61
	Dressing upper body	.80*	***.67***	.82*	.56	***.74***	.63	.62	.60	.62
	Dressing lower body	.88*	.55	.83*	.62	***.88***	***.74***	.53	.50	.49
	Toileting	.89*	.59	.83*	***.71***	***.87***	***.70***	.55	.56	.53
	Swallowing	.57*	***.49***	.61*	***.53***	***.48***	.40	***.50***	.44	.42
Sphincter care (Sp)	Bladder management	.64*	***.50***	***.64***	.67*	***.61***	.43	.42	.47	.48
	Bowel management	.65*	***.51***	***.65***	.67*	***.61***	.41	.42	.48	.50
Transfer (Tr)	Bed transfer	.91*	.54	***.86***	.65	.94*	.76	.52	.50	.47
	Toilet transfer	.91*	.56	***.87***	.68	.93*	.74	.52	.51	.50
	Shower / tub transfer	.83*	.52	***.78***	.60	.83*	***.71***	.51	.47	.45
	Car Transfer	.84*	.52	***.78***	.54	.84*	***.79***	.49	.48	.46
Loco-motion (Lo)	Walk	.64*	.43	***.60***	.33	***.65***	.71*	.43	.38	.36
	Stairs	.70*	.34	***.62***	.46	***.74***	.78*	.36	.30	.28
	Community mobility	.64*	***.70***	***.63***	.37	***.62***	.76*	.59	***.69***	***.65***
Comm-unication (Co)	Comprehension	.53	.85*	.54	.41	.42	.47	.85*	.70	***.76***
	Expression	.62	.81*	.64	.45	.51	.53	.91*	.67	.66
	Reading	.58	.78*	.62	.42	.48	.47	.85*	.65	.64
	Writing	***.60***	.69*	.63	.39	.54	.47	.78*	.58	.57
	Speech intelligibility	***.57***	.66*	.59	.37	.48	.49	.75*	.58	.51
Psycho-social adjustment (Pa)	Social interaction	.55	.84*	.58	.49	.45	.43	***.71***	.77*	***.80***
	Emotional status	.41	.63*	.43	.41	.31	.31	.48	.68*	***.61***
	Adjustm'nt to limitations	.47	.76*	.47	.41	.37	.43	.55	.79*	***.77***
	Employability	***.69***	.70*	***.69***	.41	***.63***	***.65***	***.64***	.61*	***.64***
Cognitive functions (Cf)	Problem solving	.56	.86*	.57	.43	.45	.50	.67	***.85***	.85*
	Memory	.53	.80*	.54	.47	.43	.41	.64	***.72***	.84*
	Orientation	.58	.79*	.59	.61	.47	.41	.64	***.70***	.81*
	Attention	.54	.78*	.56	.48	.44	.42	.60	***.73***	.83*
	Safety judgement	.58	.80*	.57	.43	.48	.56	.61	***.77***	.84*

[1] Corrected item-own scale / subscale correlation.
[2] Values in **bold** indicate item-other subscale correlations which do not satisfy criteria for definite scaling success.

Table 4.7

Item convergent and discriminant validity for FIM scales and subscales (N = 209)

FIM scale	No. items	Item-scale / subscale correlation		Scaling success and failure rates (%)			
		Item-own[1] range (mean)	Item-other[2] range (mean)	Definite scaling success	Probable scaling success	Probable scaling failure	Definite scaling failure
Motor	13	.63 - .92 (.79)	.30 - .66 (.50)	92.3	7.7	0	0
Cognitive	5	.77 - .85 (.80)	.50 - .60 (.54)	100	0	0	0
FIM subscale							
Self-care	6	.73 - .87 (.83)	.43 - .88 (.61)	76.6	16.7	6.7	0
Sphincter care	2	.67	.38 - .64 (.50)	60	40	0	0
Transfer	3	.82 - .94 (.90)	.44 - .88 (.62)	80	20	0	0
Locomotion	2	.68	.25 - .71 (.47)	60	30	10	0
Communication	2	.87	.35 - .77 (.52)	90	10	0	0
Social cognition	3	.76 - .84 (.80)	.29 - .72 (.49)	86.7	13.3	0	0

[1] Corrected item-total correlations for scales and subscales.
[2] Item-other correlations for scales and subscales.

Table 4.8

Item convergent and discriminant validity for FIM+FAM scales and subscales (N = 209)

FIM scale	No. items	Item-scale /subscale correlation		Scaling success and failure rates (%)			
		Item-own[1]	Item-other[2]	Definite scaling success	Probable scaling success	Probable scaling failure	Definite scaling failure
Motor	16	.57 - .91	.34 - .71	62.5	31.3	6.2	0
Cognitive	14	.63 - .86	.41 - .69	78.6	21.4	0	0
FIM subscale							
Self-care	7	.61 - .86	.43 - .88	71.4	23.8	4.8	0
Sphincter care	2	.67	.41 - .65	66.7	33.3	0	0
Transfer	4	.83 - .94	.45 - .87	75	25	0	0
Locomotion	3	.71 - .78	.28 - .74	50	50	0	0
Communication	5	.75 - .91	.37 - .76	96.7	3.3	0	0
Psychosocial adjustment	4	.61 - .79	.31 - .80	62.5	12.5	25	0
Cognitive functions	5	.81 - .85	.41 - .85	83.3	16.7	0	0

[1] Corrected item-total correlations for scales and subscales.
[2] Item-other correlations for scales and subscales.

Table 4.9

Correlations between FIM items and rotated components extracted from principal components analysis (PCA) for FIM admission scores

	PCA-1 [1]		PCA-2 [2]			
	Component		Component			
FIM item	1	2	1	2	3	4
Feeding	**.64**[3]	.36	.15	**.84**	.12	.10
Grooming	**.66**	.55	.26	**.85**	.14	.15
Bathing	**.77**	.47	**.60**	**.58**	.17	.32
Dressing upper body	**.68**	.49	.42	**.76**	.18	.14
Dressing lower body	**.87**	.27	**.71**	.45	.13	.33
Toileting	**.86**	.31	**.68**	.48	.15	.38
Bladder management	**.61**	.33	.27	.08	.16	**.76**
Bowel management	**.59**	.37	.16	.25	.06	**.78**
Bed transfer	**.91**	.23	**.76**	.38	.17	.37
Toilet transfer	**.91**	.26	**.73**	.41	.17	.38
Shower / tub transfer	**.83**	.24	**.67**	.40	.15	.36
Walk	**.67**	.16	**.67**	.20	.13	- .19
Stairs	**.80**	.02	**.84**	- .03	.09	.14
Comprehension	.21	**.86**	.08	.07	**.84**	.01
Expression	.34	**.79**	.10	.32	**.74**	- .21
Social interaction	.23	**.85**	.20	.15	**.69**	.04
Problem solving	.24	**.83**	.18	.03	**.80**	.24
Memory	.20	**.85**	.02	.03	**.75**	.33

[1] *N* = 209; based on components with eigenvalues > 1.0 and satisfying scree test.
[2] *N* = 367 (NRU database sample); based on components with eigenvalues > 1.0 and satisfying scree test.
[3] Values in **bold** indicate highest component loading for each item. When an item equally loads two or more components all these values are bolded.

Table 4.10

Correlations between items and rotated components extracted from principal components analysis (PCA) for FIM+FAM admission scores

	PCA-1 [1] (*N* = 209)				PCA-2 [2] (*n* = 105)				PCA-3 [2] (*n* = 104)			
	Components				Components				Components			
FIM+FAM item	1	2	3	4	1	2	3	4	1	2	3	4
Feeding	**.46**[3]	.09	**.50**	**.50**	.39	.15	.34	**.66**	.37	.11	**.52**	**.54**
Grooming	**.51**	.35	.41	**.47**	.47	.38	.32	**.55**	**.47**	.32	**.49**	**.45**
Bathing	**.69**	.35	.31	.32	**.68**	.36	.21	.35	**.70**	.33	.33	.30
Dressing upper body	**.59**	.35	.34	.34	**.56**	.37	.33	.32	**.58**	.31	.41	.35
Dressing lower body	**.83**	.22	.18	.27	**.80**	.32	.16	.29	**.82**	.22	.19	.27
Toileting	**.77**	.26	.17	.38	**.72**	.25	.19	.44	**.77**	.25	.22	.34
Swallowing	.22	.13	.44	**.63**	.14	.17	.27	**.77**	.23	.18	.42	**.62**
Bladder management	.40	.31	.03	**.65**	.40	.25	.03	**.69**	.40	.33	-.04	**.63**
Bowel management	.37	.31	.05	**.70**	.47	.35	-.02	**.59**	.34	.24	.03	**.74**
Bed transfers	**.85**	.18	.18	.33	**.82**	.22	.18	.37	**.83**	.14	.26	.31
Toilet transfers	**.82**	.20	.17	.38	**.80**	.25	.14	.41	**.83**	.18	.24	.33
Shower transfers	**.79**	.18	.19	.25	**.80**	.27	.10	.25	**.79**	.14	.15	.31
Car transfers	**.84**	.20	.16	.17	**.80**	.25	.19	.18	**.85**	.15	.19	.20
Walking	**.70**	.13	.26	-.02	**.69**	.08	.31	.02	**.76**	.16	.22	-.04
Stairs	**.83**	.04	.10	.06	**.84**	-.04	.16	.05	**.85**	.03	.05	.11
Community mobility	**.60**	**.55**	.27	-.15	**.55**	**.55**	.32	-.06	**.63**	.53	.24	-.18
Comprehension	.17	**.61**	**.65**	.06	.26	.51	**.73**	.02	.19	.55	**.69**	.06
Expression	.27	.44	**.77**	.12	.28	.42	**.78**	.15	.28	.37	**.80**	.11
Reading	.22	.44	**.73**	.14	.24	.39	**.78**	.18	.21	.43	**.74**	.14
Writing	.33	.32	**.70**	.11	.27	.35	**.70**	.21	.25	.30	**.74**	.09
Speech intelligibility	.26	.25	**.77**	.16	.22	.17	**.79**	.25	.33	.33	**.71**	.10
Social interaction	.16	**.73**	.39	.24	.16	**.69**	.48	.21	.14	**.74**	.41	.18
Emotional status	.07	**.67**	.15	.28	-.02	**.63**	.24	.28	.15	**.76**	.11	.25
Adjustment to limit's	.18	**.84**	.14	.07	.21	**.78**	.25	.08	.21	**.83**	.18	.04
Employability	**.56**	**.51**	.36	-.03	**.56**	**.49**	.37	.09	**.58**	.47	.38	-.06
Problem solving	.26	**.85**	.26	.04	.24	**.83**	.33	.11	.21	**.86**	.27	.01
Memory	.18	**.80**	.25	.16	.28	**.77**	.18	.17	.12	**.76**	.35	.19
Orientation	.17	**.73**	.24	.39	.19	**.75**	.18	.36	.14	**.72**	.31	.38
Attention	.20	**.79**	.18	.22	.26	**.81**	.17	.25	.17	**.79**	.21	.26
Safety judgement	.29	**.80**	.21	.10	.27	**.79**	.22	.16	.26	**.79**	.24	.17

[1] PCA-1 = total sample; based on components with eigenvalues > 1.0 and satisfying the scree test.
[2] PCA-2 and PCA-3 = samples generated by random split half.

[3] Values in **bold** indicate the highest component loading for each item. When an item equally loads two or more components all these values are bolded.

Table 4.11

Item descriptive statistics, reliability, and responsiveness for the FIM (N = 209)

FIM item	Descriptive statistics			Reliability			Responsiveness
					Reproducibility[1]		
	MEF %[2]	Floor effect %	Ceiling effect %	Item-total correlation	Intra-rater	Inter-rater	Effect size[3]
Feeding	44.5	12.0	27.3	.69	.96	.94	- .42
Grooming	36.8	12.0	36.8	.83	.92	.92	- .39
Bathing	24.9	15.3	24.9	.87	.90	.88	- .44
Dressing upper body	26.8	12.0	26.8	.82	.91	.90	- .46
Dressing lower body	28.2	28.2	12.4	.83	.95	.90	- .58
Toileting	25.8	22.5	25.8	.85	.87	.92	- .43
Bladder care	39.7	19.6	39.7	.65	.91	.85	- .30
Bowel care	46.4	13.4	46.4	.66	.89	.88	- .21
Bed transfer	25.4	17.2	19.1	.85	.95	.94	- .52
Toilet transfer	38.3	17.7	10.0	.86	.95	.96	- .46
Shower/tub transfer	29.2	29.2	1.9	.78	.91	.84	- .63
Walk	34.9	34.9	7.7	.60	.94	.84	- .69
Stairs	56.0	56.0	2.4	.62	.95	.91	- .58
Comprehension	47.4	7.2	47.4	.64	.95	.90	- .17
Expression	41.1	11.5	41.1	.70	.90	.92	- .17
Social interaction	40.2	10.0	40.2	.65	.92	.87	- .22
Problem solving	19.1	17.7	19.1	.65	.86	.89	- .18
Memory	42.6	12.4	42.6	.62	.88	.79	-. 22

[1] n = 77 for intra-rater reproducibility, and n = 89 for inter-rater reproducibility.
[2] Maximum Endorsement Frequency
[3] n = 194

Table 4.12

Item-intercorrelations for the FIM that are greater than or equal to .70 (N = 209)

Items	r[1]	Items	r
Feeding - Grooming	.78	Dressing lower body - toileting	.86
Grooming - Bathing	.79	Dressing lower body - Bed transfer	.85
Grooming - Dressing upper body	.77	Dressing lower body - Toilet transfer	.85
Grooming - Toileting	.73	Dressing lower body - Shower transfer	.79
Grooming - Bed transfer	.72	Toileting - Bed transfer	.86
Grooming - Toilet transfer	.73	Toileting - Toilet transfer	.88
Bathing - Dressing upper body	.74	Toileting - Shower transfer	.76
Bathing - Dressing lower body	.82	Bed transfer - Toilet transfer	.96
Bathing - Toileting	.79	Bed transfer - Shower transfer	.81
Bathing - Bed transfer	.79	Bed transfer - Stairs	.71
Bathing - Toilet transfer	.81	Toilet transfer - Shower transfer	.81
Bathing - Shower transfer	.75	Comprehension - Expression	.87
Dressing upper body - Dressing lower body	.75	Comprehension - Memory	.71
Dressing upper body - Toileting	.72	Social interaction - Problem solving	.75
Dressing upper body - Bed transfer	.72	Memory - Problem solving	.79
Dressing upper body - Toilet transfer	.74		

[1] Pearson's Product-Moment Correlation Coefficient

Table 4.13

Correlations between FIM-8 items and components extracted by principal components analysis (PCA)

	PCA-1[1] (*n* =209)		PCA-2 [2] (*n* =105)		PCA-3 [2] (*n* =104)	
FIM-8 item	Component 1	Component 2	Component 1	Component 2	Component 1	Component 2
Feeding	**.59** [3]	.45	.38	**.70**	.51	.51
Shower/tub transfer	**.75**	.39	.46	**.70**	**.82**	.29
Walk	**.83**	.10	.07	**.85**	**.79**	.13
Stairs	**.89**	.09	.17	**.82**	**.93**	.02
Bladder management	.44	**.63**	**.71**	.38	.52	.51
Bowel management	.42	**.67**	**.77**	.35	.50	.54
Social interaction	.12	**.85**	**.82**	.17	.09	**.88**
Memory	.07	**.85**	**.84**	.08	.08	**.87**
Eigenvalue	4.16	1.27	4.30	1.24	4.03	1.36
Variance (percent)	52.0	15.9	53.7	15.5	50.3	17.1

[1] PCA-1 = total sample; based on components with eigenvalues > 1.0 and satisfying the scree test.
[2] PCA-2 and PCA-3 = samples generated by random split half .

[3] Values in **bold** indicate the highest component loading for each item. When an item equally loads two or more components all these values are bolded.

Table 4.14

Internal consistency of four methods of scaling FIM-8 items (N = 209)

Methods of scaling FIM-8 items[1]	No. of items	Internal consistency		
		Item-total correlation	Alpha coefficient	Homogeneity coefficient
Method 1				
Total	8	.54 - .73	.86	.45
Method 2				
Motor	6	.58 - .76	.85	.50
Cognitive	2	.69	.82	.69
Method 3				
Physical	4	.56 - .72	.83	.56
Sphincter	2	.67	.80	.67
Cognitive	2	.69	.82	.69
Method 4				
Component 1	4	.56 - .72	.83	.56
Component 2	4	.61 - .65	.81	.52

[1] Methods 1, 2, and 3 are clinically-based item groupings. Method 4 is an empirically-based item grouping generated by principal components analysis.

Table 4.15

Intercorrelations between scales for methods[1] of grouping FIM-8 items (alphas in parentheses) (*N*=209)

FIM-8 scale	No. items	Method 1 Total	Method 2 Motor	Method 2 Cognitive	Method 3 Physical	Method 3 Sphincter	Method 3 Cognitive	Method 4 Component 1	Method 4 Component 2
Total	8	(.86)							
Motor	6	.96	(.85)						
Cognitive	2	.74	.51	(.82)					
Physical	4	.88			(.83)				
Sphincter	2	.82			.57	(.80)			
Cognitive	2	.74			.41	.52	(.82)		
Component 1	4	.88						(.83)	
Component 2	4	.89						.57	(.81)

[1] Methods 1, 2, and 3 are clinically-based item groupings. Method 4 is an empirically-based item grouping generated by principal components analysis.

Table 4.16

Multitrait scaling analysis of four methods of scaling FIM-8 items

Methods of scaling FIM-8 items [1]	Number of items	Item-scale correlation		Scaling success rate %	
		Item-own correlation range	Item-other correlation range	Definite	Probable
Method 1					
Total	8	.54 - .73	N/A	N/A	N/A
Method 2					
Motor	6	.58 - .76	.23 - .49	83.3	16.7
Cognitive	2	.69	.45 - .49	100	0
Method 3					
Physical	4	.56 - .72	.23 - .60	75	25
Sphincter	2	.67	.46 - .53	75	25
Cognitive	2	.69	.36 -.49	100	0
Method 4					
Component 1	4	.56 - .72	.36 - .59	50	50
Component 2	4	.61 - .65	.36 - .53	50	50

[1] Methods 1, 2, and 3 are clinically-based item groupings. Method 4 is an empirically-based item grouping generated by principal components analysis.

Table 4.17

Descriptive statistics for FIM-8 admission scores (N = 209)

FIM-8 scale	No. items	Range of scores: Possible	Range of scores: Actual	Scale mid-point	Sample mean (SD)	Floor effect %	Ceiling effect %	Skew-ness
Total	8	8 - 56	8 - 53	32	35.0 (12.2)	3.8	0	- .637
Motor	6	6 - 42	6 - 42	24	25.0 (9.6)	5.7	.5	- .356
Cognitive	2	2 - 14	2 - 14	8	9.9 (4.1)	8.1	27.8	- .689

Table 4.18

Reproducibility estimates for FIM-8 scales (N = 209)

FIM-8 scale	No. items	Reproducibility (ICC [1]) Intra-rater n = 77	Inter-rater n = 89
Total	8	.97	.96
Motor	6	.98	.96
Cognitive	2	.92	.88

[1] Intraclass Correlation Coefficient

Table 4.19

Correlations between scales of the FIM-8, FIM, and FIM+FAM (*N* = 209)

	FIM			FIM+FAM		
FIM-8 scale	Total	Motor	Cognitive	Total	Motor	Cognitive
Total	**.97**[1]	.95	.76	**.95**	.95	.80
Motor	.92	**.96**	.56	.87	**.96**	.62
Cognitive	.73	.56	**.95**	.81	.58	**.92**

[1] Values in **bold** indicate correlations between comparable scales of the FIM-8, FIM, and FIM+FAM.

Table 4.20

Intercorrelations between FIM-8 scales (alphas in parentheses) (*N* = 209)

		FIM-8 scale		
FIM-8 scale	No. items	Total	Motor	Cognitive
Total	8	(.86)		
Motor	6	.96	(.85)	
Cognitive	2	.74	.51	(.82)

Table 4.21

Correlations between the FIM-8 and other outcome measures

	Health outcome measures [1]							
	Disability				Handicap	Health status		Psychological distress
FIM-8 scale	BI[2]	MBI[3]	EDSS[4]	OPCS[5]	LHS[6]	SF-36 PCS[7]	SF-36 MCS[8]	GHQ[9]
Total	.92	.93	- .82	- .77	.25	.30	.07	- .12
Motor	.93	.94	- .84	- .80	.28	.33	.08	- .14
Cognitive	.53	.61	- .42	- .33	.05	.06	.02	- .01

[1] Sample sizes differ as not all instruments were administered at each study site. Also, the EDSS was only administered to MS patients.
[2] Barthel Index (n = 149)
[3] Modified Barthel Index (n = 60)
[4] Kurtzke Extended Disability Status Scale (n = 64)
[5] Office of Population Censuses and Surveys Overall Weighted Disability Severity Score (n = 69)
[6] London Handicap Scale (n = 121)
[7] SF-36 Physical Component Summary Score (n = 123)
[8] SF-36 Mental Component Summary Score (n = 123)
[9] General Health Questionnaire (n = 85)

Table 4.22

Correlations between FIM-8 scales and neuropsychological measures

	Neuropsychological measures [1]						
	Global decline		Reasoning			Memory	
FIM-8 scale	MMSE [2]	WAIS-VIQ [3]	HBCT [4]	WCST [5]	VESPAR [6]	CVLT [7]	RMT [8]
Total	.42	.29	- .34	.51	.48	.51	.58
Motor	.29	.23	- .32	.41	.39	.47	.53
Cognitive	.59	.36	- .26	.57	.52	.47	.47

[1] Sample sizes differ as different instruments were administered at different combinations of study sites and the VESPAR is MS specific
[2] Mini-Mental State Examination (n = 90)
[3] Wechsler Adult Intelligence Test - Verbal IQ (n = 60)
[4] Halstead Booklet Category Test (n = 44)
[5] Wisconsin Card Sorting Test (n = 40)
[6] Verbal and Spatial Reasoning Test - Total Score (n = 37)
[7] California Verbal Learning Test (n = 52)
[8] Recognition Memory Test - Visual Version (n = 49)

Table 4.23

Correlations between FIM-8 scales, age, and sex (*N* = 209)

FIM-8 scale	No. items	Age	Sex
Total	8	- .06	.06
Motor	6	- .10	.02
Cognitive	2	.06	.13

Table 4.24

Responsiveness of FIM-8 scales, and relative responsiveness compared with the FIM and FIM+FAM (n = 194)

FIM-8 scale	No. items	Mean score (SD)			Responsiveness (effect size)		
		Admission	Discharge	Change [1]	FIM-8	FIM	FIM+FAM
Total	8	35.7 (11.8)	42.5 (11.5)	- 6.8 (7.6)	- .57	- .52	- .45
Motor	6	25.5 (9.3)	31.4 (9.1)	- 5.8 (6.5)	- .63	- .58	- .57
Cognitive	2	10.1 (4.0)	11.1 (3.5)	- .94 (2.4)	- .24	- .22	- .24

[1] All change scores are statistically significant (p < .001).

Table 5.1

Responsiveness of FIM and FIM+FAM scales using five statistical methods (n = 194)

Scale	t-statistic	Relative efficiency	Effect size	Standardised response mean	Guyatt's Responsiveness Index
FIM total	- 12.1	1.0	- .52	- .87	2.59
FIM motor	- 12.3	1.03	- .58	- .88	2.90
FIM cognitive	- 6.0	.25	- .22	- .43	.69
FIM+FAM total	- 11.5	.90	- .45	- .83	2.26
FIM+FAM motor	- 12.4	1.05	- .57	- .89	2.88
FIM+FAM cognitive	- 6.9	.33	- .24	- .50	.94

Table 5.2

Comparison of rank ordering and relative responsiveness of FIM and FIM+FAM scales using five statistical methods (n = 194)

	t-statistic		Relative efficiency		Effect size		Standardised response mean		Guyatt's Responsiveness Index	
Scale	RO [1]	RR [2]	RO	RR	RO	RR	RO	RR	RO	RR
FIM total	3	1.0	3	1.0	3	1.0	3	1.0	3	1.0
FIM motor	2	1.02	2	1.03	1	1.12	2	1.01	1	1.12
FIM cognitive	6	.50	6	.24	6	.42	6	.49	6	.27
FIM+FAM total	4	.95	4	.91	4	.87	4	.95	4	.87
FIM+FAM motor	1	1.02	1	1.05	2	1.10	1	1.02	2	1.11
FIM+FAM cognitive	5	.57	5	.33	5	.57	5	.57	5	.36

[1] Rank order, 1 = most responsive instrument, 6 = least responsive instrument.
[2] Relative ratio of responsiveness compared with the FIM total scale.

Table 6.1

Comparison of disability levels in different studies

Patients / Study	Mean total score	
	FIM [1]	FIM+FAM [2]
Heterogeneous samples		
Present study	80.4	135.8
Segal *et al.* 1993 (200)	65.3	N / A [3]
Stineman *et al.* 1996 (205)	68.4	N / A
McPherson and Pentland 1996 (214)	99	154
McPherson and Pentland 1997 (216)	117	189
Stroke patients		
Present study	76.4	N / A
Granger *et al.* 1993 (211)	101.7	N / A
Brosseau *et al.* 1996 (314, 315)	72.4	N / A
Stineman *et al.* 1996 (205)	62.9	N / A
MS patients		
Present study	91.2	N / A
Granger *et al.* 1990 (208)	99.1	N / A
Brosseau 1994 (201)	88.4	N / A
Marolf *et al.* 1996 (197)	92.8	N / A

[1] Scale range 18 - 126; scale midpoint 72.
[2] Scale range 30 - 210; scale midpoint = 120.
[3] FIM+FAM not used in these studies.

Table 6.2

Comparison of methods of rating the FIM

Variable	Method of FIM rating	
	Stand alone [1]	Derived [2]
Internal consistency		
Alpha	.93	.95
Item-intercorrelations: range (mean)	.09 - .95 (.43)	.15 - .94 (.50)
Intercorrelations between scales		
total - motor	.97	.97
total - cognitive	.63	.70
motor - cognitive	.43	.51
Correlations with other variables		
Barthel Index	.93	.95
EDSS	- .78	- .84 [3]
Age	- .08	- .06

[1] Method used routinely at NRU *n* = 728.
[2] Method used in study. Patients from NRU *n* = 118.
[3] *n* = 64

Table 6.3

Mean FIM total scale change scores [1] corresponding with staff-rated changes in disability for stroke and MS patients (*n* in parentheses)

	Staff rated level of change in disability			
Diagnosis	None	Minimal	Moderate	Marked
Stroke	+ 2.33 (6)	- 7.77 (13)	- 18.38 (21)	- 36.21 (19)
MS	- 4.57 (7)	- 4.82 (22)	- 6.50 (26)	- 13.25 (4)

[1] Admission minus discharge scores. Negative change score indicates less disability on discharge relative to admission.

Appendix 1

Guidelines for rating the grooming item of the FIM and FIM+FAM

Part one: written guidelines

Grooming includes oral care, hair grooming (combing and brushing hair), washing the hands and the face, and either shaving the face or applying make-up. If there is no preference for shaving or applying make-up, then disregard. Performs safely.

No helper

7 Complete independence - subject cleans teeth or dentures, combs or brushes hair, washes the hands and the face, and either shaves the face or applies make-up, including all preparations. Performs safely.

6 Modified independence - subject requires specialised equipment including prosthesis or orthosis) to perform grooming activities, or takes more than a reasonable time, or there are safety considerations.

Helper

5 Supervision or set-up - subject requires supervision (e.g. standing by, cueing or coaxing, or set up (application of orthoses, setting out grooming equipment, and initial preparation such as applying toothpaste to brush, opening make-up containers).

4 Minimal contact assistance - subject performs $\geq$ 75% of grooming tasks.

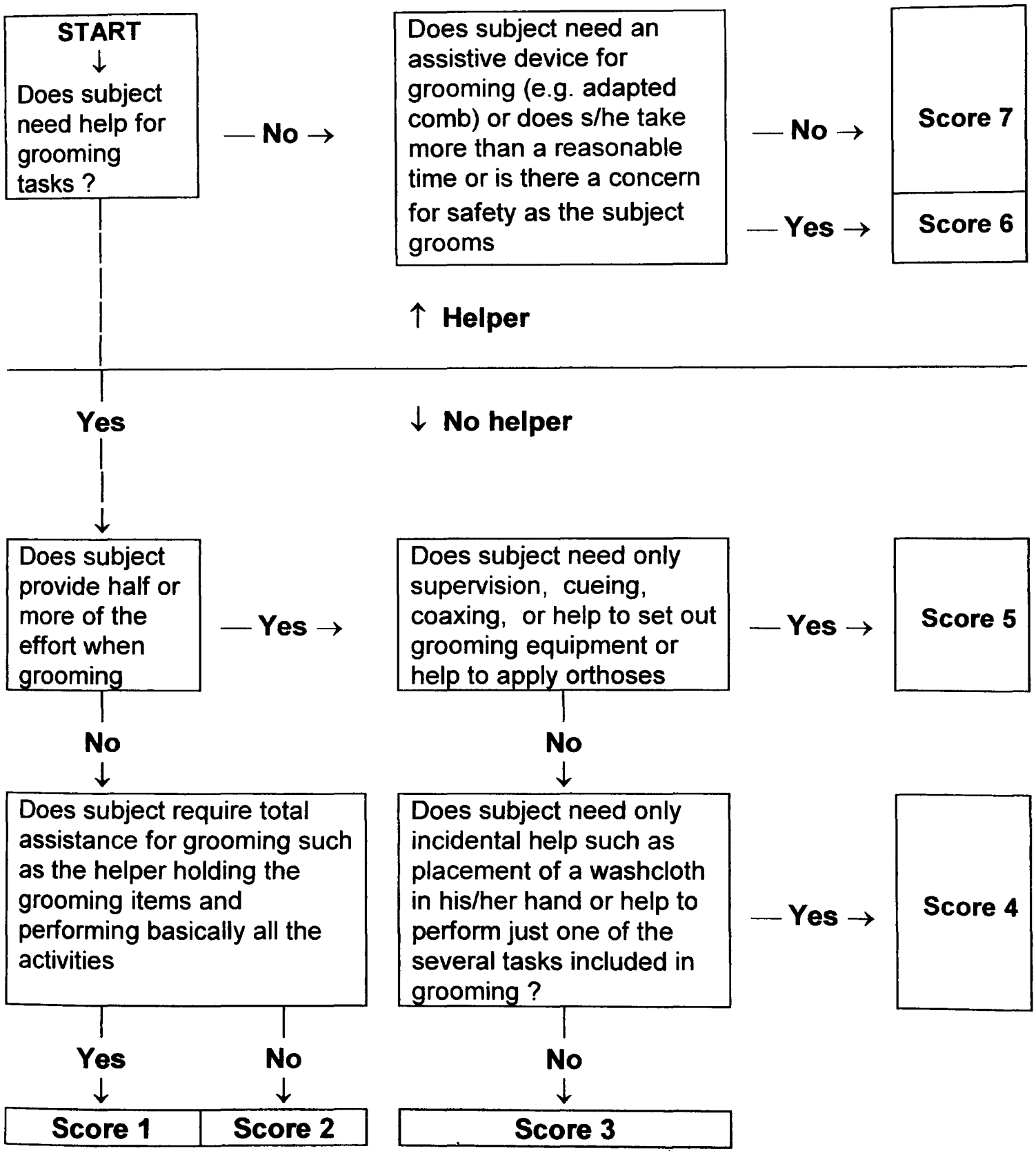
3 Moderate assistance - subject performs 50% to 74% of grooming tasks.

2 Maximal assistance - subject performs 25% to 49% of grooming tasks.

1 Total assistance - subject performs < 25% of grooming tasks.

Part two: decision tree

Grooming includes oral care, hair grooming (combing and brushing hair), washing the hands and the face, and either shaving the face or applying make-up. Note: this item may include the assessment of four or five activities depending on whether the subject chooses to shave or apply make-up. At level 7 the subject cleans his/ teeth or dentures, combs or brushes hair, washes the hands and the face, and either shaves the face or applies make-up, including all preparations. Performs independently and safely.



Appendix 2

Ethical approval, consent form, and patient information leaflet for each clinical site

THE NATIONAL HOSPITAL FOR NEUROLOGY
AND NEUROSURGERY

QUEEN SQUARE
LONDON WC1N 3BG
TEL: 071-837 3611
FAX: 071-829 8720

PATRON: Her Royal Highness The Princess of Wales
CHAIRMAN: Mrs. E. Howlett, JP
GENERAL MANAGER: A. Wheatley, CB

IFM/JAS
REF: 04/18/95

Extension 3171

19th June 1995

Dr. A. Thompson,
Consultant Neurologist,
N.H.Q.S.

Dear Dr. Thompson,

RE: DISABILITY AS AN OUTCOME OF NEUROLOGICAL REHABILITATION: COMPREHENSIVE EVALLULATION OF TEH FUNCTIONAL INDEPENDENCE MEASURE + FUNCTIONAL ASSESSMENT MEASURE (FIM + FAM) DISABILITY SCALE

I am pleased to inform you that the Joint Medical Ethics Committee approved your project at its meeting on 8th June 1995.

I would be grateful if you could ensure that the appropriate forms enclosed are completed.

With best Wishes.

Yours sincerely,

Dr. I.F. Moseley,
Chairman,
Joint Medical Ethics Committee

Encs.

RECEIVED
03 JUL 1995

R NEUROLOGY AND NEUROSURGERY

Tel: 071-837 3611

RESEARCH ON HUMAN VOLUNTEERS

Subject/Patient Consent Form

Brief description of Project: Evaluation of the Scientific properties of the Fim and Fim+FAM

Consultant(s) in charge/Director of project: Dr AJ Thompson

The subject/patient (Name): _____ Hosp. No. _____

has given his/her consent participate in the above named study.

The nature, purpose and possible consequences of the procedures involved have been explained to me by:

Name: _____

Position: _____

Signature: _____ Date _____

and Witnessed by:
Name of witness: _____

Position: _____

Address: _____

Signature: _____ Date _____

Signature Subject/Patient/Guardian: _____ Date _____

Address: _____

Please return this form to:

PATIENT SERVICES MANAGER
National Hospital for
Neurology and Neurosurgery
Queen Square
LONDON WC1N 3BG

IT IS A REQUIREMENT OF THE JOINT MEDICAL COMMITTEE THAT ANY ADVERSE EFFECTS WHICH MAY OCCUR DURING A CLINICAL TRIAL ARE REPORTED TO THE PATIENT SERVICES MANAGER IMMEDIATELY.

THE NATIONAL HOSPITAL FOR NEUROLOGY
AND NEUROSURGERY

QUEEN SQUARE
LONDON WC1N 3BG
TEL: 071-837 3611
FAX: 071-829 8720

PATRON: Her Royal Highness The Princess of Wales
CHAIRMAN: Mrs. E. Howlett, JP
GENERAL MANAGER: A. Wheatley, CB

INFORMATION FOR PATIENTS

A STUDY TO DETERMINE THE USEFULNESS OF A DISABILITY SCALE

We invite you to participate in a research project which we believe to be of potential importance. In order to help you to understand what the research is about, we are providing you with the following information which we want to be sure you understand before you agree to participate. Be sure to ask any questions you have about the information which follows and we will do our best to provide any further information you require

WHY HAVE YOU BEEN ASKED TO PARTICIPATE?

We are asking patients who are admitted to the Neuro-Rehabilitation Unit for in-patient rehabilitation on Mondays to participate.

WHAT ARE THE AIMS OF THE STUDY?

It is generally believed that rehabilitation is important for people with neurological disease, but this has yet to be proven. To do this staff at rehabilitation units must be able to measure the effect of rehabilitation accurately. This study will evaluate a measure to determine if it is accurate enough to be used to assess the effect of in-patient rehabilitation. This is achieved by collecting information from you and the staff who look after and treat you during your stay at the Rehabilitation Unit.

HOW DOES THIS INVOLVE YOU?

If you agree to paticipate we would ask you to answer some questions and fill in 2 questionnaires within a few days of admission to the Neuro-Rehabilitation Unit and again before discharge. This takes approximately 20 minutes. We will also ask you to undertake some tests to assess your memory. This will take about 2 hours. None of these tests will interfere with your rehabilitation in any way. The study will result in NO discomforts or hazards NOR extra visits to hospital than would ordinarily be the case. It will not interfere or affect any other medical problems you may have. The research will not be of special benefit to you during your rehabilitation. All answers are in confidence and will be coded so that they are not identifiable by name.

You are completely free not to participate and may withdraw from the study at any time. This will not jeopardise the ordinary course of medical treatment (or course of study, if you are a student volunteer). This will not affect your rights at all. You understand that in the event of injury caused by your participation in research, you will be compensated irrespective of the negligence of the researchers.

If you have any questions please do not hesitate to ask. More information can be obtained from Dr. Jeremy Hobart (Research Doctor), Jenny Freeman (Research Therapist) or Dr. Thompson, Consultant Neurologist at the Rehabilitation Unit

Contact telephone number: 0171 837 3611 ext. 3341

THANK YOU FOR YOUR CO-OPERATION

ALL CORRESPONDENCE TO BE ADDRESSED TO:
MR MARK KENDALL
ELCHA RESEARCH ETHICS COMMITTEE
61 PHILPOT STREET
WHITECHAPEL
LONDON, E1 2JH
TEL: 071-377-7325

EAST LONDON & THE CITY
HEALTH AUTHORITY

Chairman
Professor Frances Heidensohn

Dr R Greenwood
Regional Neuro-Rehabilitation Unit
Homerton Hospital
Homerton Way
London
E9 6SR

Our ref: MS/MK/cat

25 April 1995

Dear Dr Greenwood

Re: P/95/61 - Disability as an outcome of neurological rehabilitation: comprehensive evaluation of the functional independence measure + functional assessment measure (FIM + FAM) disability scale

Further to your letter of 7 April addressing the concerns of the Research Ethics Committee, I now have pleasure in taking Chairman's Action in accepting the above study as ethically satisfactory and will report this to the next full meeting of the Ethics Committee.

Please note the following conditions to the approval:

1. The Committee's approval is for the length of time specified in your application. If you expect your project to take longer to complete (ie collection of data), a letter from the principal investigator to the Chairman will be required to further extend the research. This will help the Committee to maintain comprehensive records.

2. Any changes to the protocol must be notified to the Committee. Such changes may not be implemented without the Committee or Chairman's approval.

3. The Committee should be notified immediately of any serious adverse events or if the study is terminated prematurely.

4. You are responsible for consulting with colleagues and/or other groups who may be involved or affected by the research, such as extra work for laboratories.

/Continued

5. You must ensure that, where appropriate, nursing and other staff are made aware that research in progress on patients with whom they are concerned has been approved by the Committee.

6. The Committee should be sent one copy of any publication arising from your study, or a summary if there is to be no publication.

Please quote the above study number in any future related correspondence.

Yours sincerely

M SWASH MD FRCP FRCPath
Chairman
ELCHA Research Ethics Committee

REGIONAL NEURO-REHABILITATION UNIT, HOMERTON HOSPITAL

WRITTEN CONSENT FORM

Title of research proposal:

Disability as an Outcome of Neurological Rehabilitation: Comprehensive Evaluation of the Functional Independence Measure + Functional Assessment Measure (FIM + FAM) Disability Scale

E.C. No.
Name of Patient:
Address:

I have read the attached information on the research in which I have been asked to participate and have been given a copy to keep. I have had the opportunity to discuss the details and ask questions about this information. The Investigator has explained the nature and purpose of the research and I believe that I understand what is being proposed. For example, I understand that this trial is part of a research project designed to promote medical knowledge, and that it has been approved by the East London & City Health Authority Research Ethics Committee. I have been informed that the proposed study involves monitoring and special examinations which have been explained to me, together with possible risk involved. I understand that my personal involvement and my particular data from this trial will remain strictly confidential. Only researchers involved in the trial will have access, or where applicable, the industrial sponsor which funded the research. I also understand that, where appropriate, my General Practitioner will be informed that I have taken part in this study. I hereby fully and freely consent to participate in the study which has been fully explained to me.

In circumstances where a patient is deemed unable to give informed consent, a relative may give their permission for their involvement in the study.

PATIENT'SNAME:(BLOCK CAPITALS).....

PATIENT'S NAME:SIGNATURE.....

WITNESS' NAME:.....

WITNESS' SIGNATURE: .....

IS A RELATIVE OF THE ABOVE, I GIVE MY PERMISSION FOR THEIR INVOLVEMENT IN THIS STUDY AS THEY ARE UNABLE TO GIVE INFORMED CONSENT THEMSELVES :

NAME OF RELATIVE:.....

PATIENT'S RELATIVE NAME:SIGNATURE.....

INVESTIGATOR'S NAME: .....

INVESTIGATOR'S SIGNATURE: .....

DATE:.....

The following should be signed by the Clinician/Investigator responsible for obtaining consent.

As the Clinician/Investigator responsible for this research or a designated deputy, I confirm that I have explained to the patient/volunteer named above the nature and purpose of the research to be undertaken.

CLINICIAN'S NAME: .....

CLINICIAN'S SIGNATURE: .....

DATE: .....

Subjects are warned not to take part in more than one study at any time.

If you are at all concerned about this trial or note any untoward effect of any drug you are receiving, please contact:

Dr. RICHARD GREENWOOD

Tel. No. 081 919 7970.....

REGIONAL NEURO-REHABILITATION UNIT, HOMERTON HOSPITAL

"INVITATION TO PARTICIPATE IN A RESEARCH PROJECT"

INFORMATION FOR PATIENTS

We invite you to participate in a research project which we believe to be of potential importance. In order to help you to understand what the research is about, we are providing you with the following information which we want to be sure you understand before you formally agree to participate. Be sure to ask any questions you have about the information which follows and we will do our best to explain and to provide any further information you require.

TITLE OF RESEARCH PROJECT

Disability as an Outcome of Neurological Rehabilitation: Comprehensive Evaluation of the Functional Independence Measure + Functional Assessment Measure (FIM + FAM) Disability Scale

WHY HAVE YOU BEEN ASKED TO PARTICIPATE?

We are asking patients who are admitted to the Regional Neuro-Rehabilitation Unit at the Homerton Hospital for in-patient rehabilitation on a Monday to participate in the study.

WHAT ARE THE AIMS OF THE STUDY?

It is generally believed that rehabilitation is important for people with neurological disease, but this has yet to be proven. In order to answer this question it is necessary to have an accurate way of measuring the effects. This study will evaluate a measure to determine if it is dependable enough to be used to assess the effect of in-patient rehabilitation.

HOW DOES THIS INVOLVE YOU

We would like you to answer some questions and fill in 2 questionnaires within a few days of admission to the Neuro-Rehabilitation Unit and again before discharge. This takes approximately 30 minutes, and will not interfere with your rehabilitation in any way. All answers are in confidence and will be coded so that they are not identifiable by name. You are completely free not to agree to participation and may withdraw from the study at any time.

You understand that in the event of injury caused by your participation in the research, you will be compensated irrespective of the negligence of the researchers.

General information on patients' rights, particularly as regards participation in research studies may also be obtained from my local Community Health Council.

More information can be obtained from Dr. Greenwood, Consultant Neurologist at the Rehabilitation Unit

24 hour contact number: 081 919 7969

REGIONAL NEURO-REHABILITATION UNIT, HOMERTON HOSPITAL

"INVITATION TO PARTICIPATE IN A RESEARCH PROJECT"

INFORMATION FOR RELATIVES

We invite your relative to participate in a research project which we believe to be of potential importance. In order to help you to understand what the research is about, we are providing you with the following information which we want to be sure you understand before you formally give permission for your to relative participate. Be sure to ask any questions you have about the information which follows and we will do our best to explain and to provide any further information you require. We feel that your relative is currently unable to give informed consent and therefore we are asking for your permission (relatives cannot give consent in the place of patients).

TITLE OF RESEARCH PROJECT

Disability as an Outcome of Neurological Rehabilitation: Comprehensive Evaluation of the Functional Independence Measure + Functional Assessment Measure (FIM + FAM) Disability Scale

WHY HAS YOUR RELATIVE BEEN ASKED TO PARTICIPATE?

We are asking patients who are admitted to the Regional Neuro-Rehabilitation Unit at the Homerton Hospital for in-patient rehabilitation on a Monday to participate.

WHAT ARE THE AIMS OF THE STUDY?

It is generally believed that rehabilitation is important for people with neurological disease, but this has yet to be proven. It is generally believed that rehabilitation is important for people with neurological disease, but this has yet to be proven. In order to answer this question it is necessary to have an accurate way of measuring the effects. This study will evaluate a measure to determine if it is dependable enough to be used to assess the effect of in-patient rehabilitation.

HOW DOES THIS INVOLVE YOU

If you give permission to their paticipation we would ask you to answer some questions and fill in 2 questionnaires within a few days of admission to the Neuro-Rehabilitation Unit and again before discharge. This takes approximately 30 minutes. All answers are in confidence and will be coded so that they are not identifiable by name. You are completely free not to agree to the participation of your relative.

General information on patients' rights, particularly as regards particiipation in research studies may also be obtained from my local Community Health Council.

More information can be obtained from Dr. Greenwood, Consultant Neurologist at the Rehabilitation Unit

24 hour contact number: 081 919 7969

17 MAR 1995 380

SOUTHAMPTON AND SOUTH WEST HAMPSHIRE HEALTH COMMISSION

Oakley Road, Southampton SO16 4GX
Telephone: (01703) 725400, Fax: (01703) 725457

TEW/3/WGP/JS

COPY
for Jeremy

14 March 1995

Professor D L McLellan
Europe Professor of Rehabilitation
Research Unit
University Faculty of Medicine
Southampton General Hospital

Dear Professor McLellan

Submission No. 65/95 - Disability as an Outcome of Neurological Rehabilitation: Comprehensive Evaluation of the Functional Independence Measure + Functional Assessment measure (FIM + FAM) Disability Scale

The Joint Ethics Committee considered the above application at its recent meeting. I am pleased to inform you that ethical approval was given to this study.

Would you please ensure that a record is made in the Medical Records of patients who agree to participate in a research project, to the effect that they have given their consent to involvement in this research study. The title of the research project should be clearly indicated. Please note that this applies only to patients who do wish to be involved in a study and not for patients who do not wish to participate.

Should any unforeseen problem of either an ethical or procedural nature arise during the course of this research where you feel that the Joint Ethics Committee may be of assistance, please do not hesitate to contact me.

I would be grateful if you could complete the enclosed questionnaire, and forward it to Ms Frances Marsden, Deputy Director of Administration Services, University of Southampton. This is necessary to adhere to University procedures on insurance cover.

Yours sincerely

Dr. T. E. Woodcock
Honorary Secretary
Joint Ethics Committee

Enc.

REHABILITATION RESEARCH UNIT SOUTHAMPTON GENERAL HOSPITAL

RESEARCH ON HUMAN VOLUNTEERS

Subject / Patient Consent Form

rief description of project:

Study to investigate the usefulness of the FIM+FAM disability scale

onsultant in charge of project: **Professor DL McLellan**

ubject / Patient Name:_____Hospital number:_____

as given her/his consent to participate in the above study. The nature, purpose and possible onsequences of the procedures involved have been fully explained to me by:

Name:_____

Position:

Signature:_____Date:_____

nd witnessed by:_____

Name of witness:_____

Position/Address:_____

Signature:_____Date:_____

Signature of patient/subject/guardian:_____Date:_____

Address: _____

_____

INFORMATION FOR PATIENTS

A STUDY TO DETERMINE THE USEFULNESS OF A DISABILITY SCALE

/e invite you to participate in a research project which we believe to be of potential importance. ı order to help you to understand what the research is about, we are providing you with the)llowing information which we want to be sure you understand before you formally agree to articipate. Be sure to ask any questions you have about the information which follows and we 'ill do our best to explain and to provide any further information you require.

/HY HAVE YOU BEEN ASKED TO PARTICIPATE?

/e are asking all patients with neurological problems who are admitted to the Research ehabilitation Unit at Southampton General Hospital for in-patient rehabilitation to participate.

/HAT ARE THE AIMS OF THE STUDY?

: is generally believed that rehabilitation is important for people with neurological disease, but ıis has yet to be proven. To do this staff at rehabilitation units must be able to measure the effect f rehabilitation accurately. This study will evaluate a measure to determine if it is accurate nough to be used to assess the effect of in-patient rehabilitation. This is achieved by collecting ıformation from you and the staff who look after and treat you during your stay at the .ehabilitation Unit.

IOW DOES THIS INVOLVE YOU

you agree to participate we would ask you to answer some questions and fill in 2 questionnaires 'ithin a few days of admission to the Rehabilitation Unit and again before discharge. This takes pproximately 20 minutes. None of these tests will interfere with your rehabilitation in any way. he study will result in NO discomforts or hazards NOR extra visits to hospital than would rdinarily be the case. It will not interfere or affect any other medical problems you may have. he research will not be of special benefit to you during your rehabilitation. All answers are in onfidence and will be coded so that they are not identifiable by name.

PTO

ou are completely free not to participate and may withdraw from the study at any time. This ill not jeopardise the ordinary course of medical treatment (or course of study, if you are a udent volunteer). This will not affect your rights at all. You understand that in the event of ijury caused by your participation in the research, you will be compensated irrespective of the egligence of the researchers.

you have any questions please do not hesitate to ask. More information can be obtained from rofessor McLellan, Consultant Neurologist at the Rehabilitation Unit

THANK YOU FOR YOUR CO-OPERATION

ontact telephone number: 0703 796466

Appendix 3

Barthel Index (BI) [1]

1 Feeding

0 = unable;

1 = needs help cutting, spreading butter, etc.;

2 = independent

2 Transfer

0 = unable, no sitting balance;

1 = major help (one or two people, physical), can sit;

2 = minor help (verbal or physical);

3 = independent

3 Grooming

0 = needs help with personal care;

1 = independent face / hair / teeth / shaving (implements provided)

4 Toilet use

0 = dependent;

1 = needs some help but can do something alone;

2 = independent (on and off, dressing, wiping)

5 Bathing

0 = dependent;

1 = independent (or in shower)

---

[1] Wade and Collin 1988 (226)

6 Mobility

0 = immobile;

1 = wheelchair independent, including corners;

2 = walks with the help of one person (verbal or physical);

3 = independent (but may use aid: for example, stick)

7 Stairs

0 = unable;

1 = needs help (verbal, physical, carrying aid);

2 = independent

8 Dressing

0 = dependent;

1 = needs help but can do about half unaided;

2 = independent (including buttons, zips, laces, etc.)

9 Bowel

0 = incontinent (or needs to be given enemata);

1 = occasional accident (once a week);

2 = continent

10 Bladder

0 = incontinent, or catheterised and unable to manage alone;

1 = occasional accident (once a week);

2 = continent

Appendix 4

Modified Barthel Index (MBI) [1]

A score of zero is given in all of the below activities when the patient cannot meet the criteria as defined

1 Feeding

2 = independent. The patient can feed him/herself a meal from a tray or table when someone puts the food within reach. He/she must put on an assistive device if this is needed, cut up the food, use salt and pepper, spread butter etc. He/she must accomplish this in a reasonable time. Food should not be pureed, soft or cut up, the patient should require no supervision and must be aware of the need to eat and appropriateness of time and place.

1 = some help is necessary (with cutting up food etc., as listed above) or encouragement to commence eating but the patient is then able to feed him/herself without further assistance or supervision.

2 Moving from wheelchair to bed and return

3 = independent in all phases of this activity. Patient can safely approach the bed in his/her wheel chair, lock brakes, lift footrests, move safely to bed, lie down, come to a sitting position on the side of the bed, change the

---

[1] Novick *et al.* 1996 (237)

position of the wheel chair, if necessary, to transfer back into it safely and return to the wheelchair. Patient must recognise the need to transfer and do so in appropriate circumstances without supervision.

2 = either some minimal help is needed in some step of this activity or the patient needs to be reminded or supervised for safety of one or more parts of this activity.

1 = patient can come to a sitting position without the help of a second person but needs to be lifted out of bed, or if s/he transfers with a great deal of help, e.g. one strong / skilled or two normal persons.

3 Doing personal toilet (grooming)

1 = patient can wash hands and face, comb hair, clean teeth, and shave. He may use any kind of razor but must put in blade or plug without help as well as get it from draw or cabinet. Female patients must put on own make-up if use, but need not braid or style hair. The patient must recognise the need to groom and be able to request toiletries as required. A patient who requires telling to wash/shave is dependent.

0 = patient requires any degree of assistance, physical or verbal.

4 Getting on and off toilet

2 = patient must be able to get on and toilet, fasten and unfasten clothes, prevent soiling of clothes and use toilet paper without help. S/he may use a wall bar or other stable object for support if required. if it is necessary to use

the bedpan instead of a toilet, he/she must be able to place it on a chair, empty it and clean it. Patient must both recognise the need to use the toilet and be able to get there independently.

1 = patient needs help because of imbalance or in handling clothes or in using toilet paper or requires directing or moving to the toilet.

5 Bathing self

1 = patient may use a bath tub, shower, or take a complete sponge bath. S/he must be able to do all the steps involved in whatever method is employed without another person being present. The patient recognised that he/she needs a bath/shower.

6a Walking on a level surface

3 = patient can walk at least 50 yards on the level without help or supervision. S/he may wear braces or prostheses and use crutches, canes or a walkerette but not a rolling walker. S/he must be able to lock and unlock brakes if used, assume the standing position and sit down, get the necessary mechanical aids into position for use, and dispose of them when sitting. (Putting on and taking off braces is scored under dressing.) S/he must be able to negotiate obstacles in home or ward environment. Walking must be purposeful and constructive.

2 = patient needs help or supervision with any of the above but can walk at least 50 yards with a little help.

6b Propelling a wheelchair

1 = if patient cannot ambulate but can propel a wheelchair independently. S/he must be able to go round corner, turn round, manoeuvre the chair to a table, bed, toilet etc. S/he must be able to push a chair 50 yards. Do not score this item if patients gets scored for walking. Movement must be purposeful and constructive and not demand constant supervision or restraint.

7 Ascending or descending stairs

2 = patient is able to go up and down a flight of stairs without supervision or help. S/he may use handrails, crutches, or canes when needed. S/he must be able to carry canes or crutches as s/he ascends or descends stairs. Climbing of stairs must be purposeful, e.g. a person who needs to told to go upstairs to bed / toilet etc. needs supervision

1 = patient needs help or supervision with any of the above items.

0 = patient unable to climb stairs or mobility on stairs, because of cognitive impairment, demands constant supervision.

8 Dressing and undressing

2 = patient is able to put and remove and fasten all clothing and tie shoelaces (unless it is necessary to use adaptations for this). The activity includes putting on and removing and fastening corset, braces or artificial limbs etc. when these are prescribed, but not bras. Such special clothing as suspenders, loafer shoes, dresses that open down the front may be

used when necessary. Patient must recognise the need to dress and undress and do so at appropriate times.

1 = patients needs help in putting on and removing or fastening any clothing. S/he must be able to do at least half the work her/himself. S/hr must accomplish this in a reasonable time. S/he requires to be instructed to dress or undress, or requires help with selecting clothes.

9 Controlling bowels

2 = patient is able to control bowels and have no accidents. S/he can use a suppository or take an enema when necessary (as for spinal cord patients who have had bowel training). Patient requires no staff supervision whatsoever to avoid accidents and does not defecate in inappropriate places

1 = patient requires help in using a suppository or enema, or has occasional accidents (occasional = one a week), or requires supervision of staff to avoid accidents.

0 = incontinent or frequently defecates in inappropriate places.

10 Controlling bladder

2 = patient is able to control bladder day and night. Spinal cord injury patients (or other) who use an external device (or catheter) must put them on independently, clean and empty leg bag, and stay dry day and night.

1 = patient has occasional accidents, cannot wait for the bedpan, get to the toilet in time, needs help with an external device. requires staff supervision

to avoid accidents (e.g. to be woken during the night) or occasionally micturates in inappropriate places.

0 = incontinent or frequently micturates in inappropriate places

Appendix 5

Kurtzke Expanded Disability Status Scale (EDSS) [1]

0 Normal neurological exam (all grade 0 in Functional Systems [FS]; Cerebral grade 1 acceptable.)

1.0 No disability, minimal signs in one FS (i.e. grade 1 excluding Cerebral grade 1.)

1.5 No disability, minimal signs in one or more FS (more than one grade 1 excluding cerebral grade 1.)

2.0 Minimal disability in one FS (one FS grade 2, others 0 or 1.)

2.5 Minimal disability in two FS (two FS grade 2, others 0 or 1.)

3.0 Moderate disability in one FS (one FS grade 3, others 0 or 1), or mild disability in three or four FS (three/four FS grade 2, other 0 or 1) though fully ambulatory.

3.5 Fully ambulatory but with moderate disability in one FS (one grade 3) and one or two FS grade 2; or two FS grade 3; or five FS grade 2 (other 0 or 1.)

4.0 Fully ambulatory without aid, self-sufficient, up and about some 12 hours a day despite relatively severe disability consisting of one FS grade 4 (others 0 or 1), or combinations of lesser grades exceeding limits of previous steps. Able to walk without aid or rest some 500 meters.

---

[1] Kurtzke 1983 (38)

4.5 Fully ambulatory without aid, up and about much of the day, may otherwise have some limitations of full activity or require minimal assistance; characterised by relatively severe disability, usually consisting of one FS grade 4 (others 0 or 1) or combinations of lesser grades exceeding limits of previous steps. Able to walk without aid or rest for some 300 meters.

5.0 Ambulatory without aid or rest for about 200 meters; disability severe enough to impair full daily activities (e.g. to work full day without special provisions). (Usual FS equivalents are one grade 5 alone, other 0 or 1; or combinations of lesser grades usually exceeding specifications for step 4.0.)

5.5 Ambulatory without aid or rest for about 100 meters; disability severe enough to preclude full daily activities. (Usual FS equivalents are one grade 5 alone, other 0 or 1; or combinations of lesser grades usually exceeding specifications for step 4.0.)

6.0 Intermittent or unilateral constant assistance (cane, crutch, or brace) required to walk about 100 meters without resting. (Usual FS equivalents are combinations with more than two FS grade 3+.)

6.5 Constant bilateral assistance required to walk about 20 meters without resting. (Usual FS equivalents are combinations with more than two FS grade 3+.)

7.0 Unable to walk beyond five meters even with aid, essentially restricted to wheelchair; wheels self in standard wheelchair and transfers alone; up and about in wheelchair some 12 hours a day. (Usual FS equivalents are combinations with more than two FS grade 4+; very rarely, pyramidal grade 5 alone.)

7.5 Unable to take more than a few steps; restricted to wheelchair; may need aid in transfer; wheels self but cannot carry on in standard wheel chair a full day; may require motorized wheelchair. (Usual FS equivalents are combinations with more than two FS grade 4+.)

8.0 Essentially restricted to bed or chair or perambulated in wheelchair; but may be out of bed itself much of the day; retains many self-care functions; generally has effective use of arms. (Usual FS equivalents are combinations, generally grade 4+ in several systems.)

8.5 Essentially restricted to bed much of the day; has some effective use of arm(s); retains some self-care functions. (Usual FS equivalents are combinations, generally grade 4+ in several systems.)

9.0 Helpless bed patient; can communicate and eat. (Usual FS equivalents are combinations, most grade 4+.)

9.5 Totally helpless bed patient; unable to communicate effectively or eat/swallow. (Usual FS equivalents are combinations, almost all grade 4+.)

10 Death due to MS

Appendix 6

Office of Population Censuses and Surveys Disability Scales (OPCS) [1]

1 Dimensions of disability

1.1 Locomotion

Code	Description	Severity score
L1	Cannot walk at all	11.5
L2	Can only walk a few steps without stopping or severe discomfort; cannot walk up and down one step	9.5
L3	Has fallen 12 or more times in the last year	7.5
L4	Always needs to hold on to something to keep balance	7.0
L5	Cannot walk up and down a flight of 12 stairs	6.5
L6	Cannot walk 50 yards (metres) without stopping or severe discomfort	5.5
L7	Cannot bend down far enough to touch knees and straighten up again	4.5
L8	Cannot bend down and pick something up from the floor and straighten up again	4.0
L9	Cannot walk 200 yards (metres) without stopping or severe discomfort; or can only walk up and down a flight of 12 stairs if holds on and takes a rest; or often needs to hold on to something to keep balance; or has fallen three or more times in the last year	3.0
L10	Can only walk up and down a flight of 12 stairs if holding on (doesn't need a rest).	2.5
L11	Cannot bend down to sweep up something from the floor and straighten up again	2.0
L12	Can only walk up and down a flight of stairs if going side- ways or one step at a time	1.5
L13	Cannot walk 400 yards (metres) without stopping or severe discomfort	0.5

1.2 Eating, drinking, and digestion

Code	Description	Severity score
EDD1	Suffers from problems with eating, drinking, or digestion which severely affects the ability to lead a normal life	0.5

[1] Martin *et al.* 1988 (175)

1.3 **Disfigurement** (scars, blemishes, and deformities)

Code	Description	Severity score
DF1	Suffers from a scar, blemish, or deformity which severely affects the ability to lead a normal life	0.5

1.4 Reaching and stretching

Code	Description	Severity score
RS1	Cannot hold out either arm in front to shake hands	9.5
RS2	Cannot put either arm up to the head to put a hat on	9.0
RS3	Cannot put either hand behind the back to put a jacket on or to tuck his shirt in	8.0
RS4	Cannot raise either arm above the head to reach for some-thing	7.0
RS5	Has difficulty holding either arm in front to shake hands with someone	6.5
RS6	Has difficulty putting either arm to his or her head to put a hat on	5.5
RS7	Has difficulty putting either hand behind the back to put a jacket on or to tuck his shirt in	4.5
RS8	Has difficulty raising either arm above the head to reach for something	3.5
RS9	Cannot hold one arm out in front or up to the head (but can with the other arm)	2.5
RS10	Cannot put one arm behind the back to put on a jacket or to tuck his shirt in (but can with the other arm); or has difficulty putting one arm behind the back to put a jacket on, or to tuck his shirt in; or putting one arm out in front, or up to the head (but no difficulty with the other arm)	1.0

1.5 Personal care

Code	Description	Severity score
PC1	Cannot feed self without help; or cannot go to and use the toilet without help	11.0
PC2	Cannot get into and out of bed without help; or cannot get into and out of a chair without help	9.5
PC3	Cannot wash hands and face without help; or cannot dress and undress without help	7.0
PC4	Cannot wash all over without help	4.5
PC5	Has difficulty feeding self; or has difficulty getting to and using the toilet	2.5
PC6	Has difficulty getting in and out of bed; or has difficulty getting in and out of a chair	1.0

1.6 Dexterity

Code	Description	Severity score
D1	Cannot pick up and hold a mug of coffee with either hand	10.5
D2	Cannot turn a tap (faucet) or control knobs on a cooker with either hand	9.5
D3	Cannot pick up and carry a pint of milk or squeeze the water from a sponge with either hand	8.0
D4	Cannot pick up a small object such as a safety pin with either hand	7.0
D5	Has difficulty picking up and pouring from a full kettle, or serving food from a pan using a spoon or ladle	6.5
D6	Has difficulty unscrewing the lid of a coffee jar or using a pen or pencil	5.5
D7	Cannot pick up and carry a 5 lb. (2 kg) bag of potatoes with either hand	4.0
D8	Has difficulty in wringing out light washing or using a pair of scissors	3.0
D9	Can pick up and hold a mug of tea or coffee with one hand but not with the other	2.0
D10	Can turn a tap or control a knob with one hand but not with the other; or can squeeze the water from a sponge with one hand but not with the other	1.5
D11	Can pick up a small object such as a safety pin with one hand but not with the other, or can pick up and carry a pint of milk with one hand but not with the other; or has difficulty in tying a bow in laces or strings	0.5

1.7 Continence

Code	Description	Severity score
CO1	No voluntary control over bowels	11.5
CO2	No voluntary control over bladder	10.5
CO3	Loses control of bowels at least once every 24 hours	10.0
CO4	Loses control of bladder at least once every 24 hours	8.0
CO5	Loses control of bowels at least once a week	8.0
CO6	Loses control of bowels at least twice a month	6.5
CO7	Loses control of bladder at least once a week	5.5
CO8	Loses control of bowels at least once a month	5.0
CO9	Loses control of bladder at least twice a month; or loses control of bowels occasionally	4.0
CO10	Loses control of bladder at least once a month	2.5
CO11	Loses control of bladder occasionally; or uses a device to control bowels or bladder	1.0

1.8 Communication

Code	Description	Severity score
C1	Is impossible for people who knew him/her well to under- stand; or finds it impossible to understand people who know him/ her well	12.0
C2	Is impossible for strangers to understand; or is very difficult for people who know him/her well to understand; or finds it impossible to understand strangers; or finds it very difficult to understand people who know him/her well	8.5
C3	is very difficult for strangers to understand; or is quite difficult for people who know him/her well to understand; or finds it difficult to understand strangers; or finds it quite difficult to understand people who know him/her well	5.5
C4	Is quite difficult for strangers to understand; or finds it quite difficult to understand strangers	2.0
C5	Other people have some difficulty in understanding him/her; or has some difficulty understanding what other people say or what they mean	1.0

1.9 Behaviour

Code	Description	Severity score
B1	Gets so upset that he or she hits other people, or injures him/herself.	10.5
B2	Gets so upset that he or she breaks or rips up things	7.5
B3	Feels the need to have someone present all the time	7.0
B4	Finds relationships with members of the family very difficult	6.0
B5	Often has outbursts of temper at other people with very little cause	4.0
B6	Finds relationships with people outside the family very difficult	2.5
B7	Sometimes sits for hours doing nothing	1.5
B8	Finds it difficult to stir him/herself to do things; or often feels aggressive or hostile towards other people	0.5

1.10 Intellectual functioning

Code	No. of problems	Severity score	Number of problems from the following list:
I1	11	13.0	Often forgets what he or she was supposed to be doing in the middle of something
I2	10	12.0	Often loses track of what is being said in the middle of a conversation
I3	9	10.5	Thoughts tend to be muddled and slow
I4	8	9.5	Often gets confused about what time of day it is
I5	7	8.0	Cannot watch a half-hour TV programme all the way through and tell someone what it was about
I6	6	7.0	Cannot remember and pass on a message correctly
I7	5	6.0	Often forgets to turn off things such as fires, cookers, or taps (faucets)
I8	4	4.5	Often forgets names of people in the family or friends seen regularly
I9	3	3.5	Cannot read a short article in newspaper
I10	2	2.0	Cannot write a short letter to someone without help
I11	1	1.0	Cannot count well enough to handle money

1.11 Seeing

Code	Description	Severity score
S1	Cannot tell by the light where the windows are	12.0
S2	Cannot see the shapes of furniture in a room	11.0
S3	Cannot see well enough to recognise a friend if close to his face	10.0
S4	Cannot see well enough to recognise a friend who is an arm's length away	8.0
S5	Cannot see well enough to read a newspaper headline	5.5
S6	Cannot see well enough to read a large print book	5.0
S7	Cannot see well enough to recognise a friend across a room	4.5
S8	Cannot see well enough to recognise a friend across a road	1.5
S9	Has difficulty seeing to read ordinary newspaper print	0.5

1.12 Consciousness

Code	'Fit' score	Severity score	Add scores for the following items relating to epileptic fits:
CS1	13.8	12.5	*Frequency of fits:*
			0 = less than once a year
CS2	12.8 - 13.0	11.5	1 = once a year but fewer than four times a year
CS3	11.8	10.5	2 = four times/year but less than once a month
CS4	10.8	10.0	3 = once a month but less than once a week
CS5	9.8 - 10.0	9.0	4 = once a week but less than every day
CS6	8.8 - 9.0	8.0	5 = every day
CS7	7.8 - 8.0	7.0	*Timing of fits:*
CS8	6.8 - 7.0	6.0	1 = only has fits during the night
CS9	5.8 - 6.0	5.0	3.8 = only has fits at night or on awakening
CS10	4.8 - 5.0	4.0	5.8 = only has fits at night, on awakening or in the evening
CS11	4.0	3.0	6.8 = has fits during the daytime
CS12	3.0	2.0	*Warning of fit:*
CS13	2.0	1.0	0 = always has a warning before a fit
			1 = has fits without warning
CS14	1.0	0.5	*Consciousness in fit:*
			0 = does not lose consciousness
			1 = loses consciousness during fit

1.13 Hearing

Code	Description	Severity score
H1	Cannot hear at all	11.0
H2	Cannot follow a TV programme with the volume turned up	8.5
H3	Has difficulty hearing someone talking in a loud voice in a quiet room	6.0
H4	Cannot hear a doorbell, alarm clock, or telephone bell	5.5
H5	Cannot use the telephone	4.0
H6	Cannot follow a TV programme at a volume others find acceptable	2.0
H7	Difficulty hearing someone talking in a normal voice in a quiet room	1.5
H8	Difficulty in following a conversation against background noise	0.5

2 Overall weighted disability severity score (WSS)

This is computed from ten of the 13 OPCS disability scales (eating, drinking, digestion, disfigurement, and consciousness are excluded) using the following formula:

WSS = worst score + (0.4 x second worst score) + (0.3 x third worst score)

3 **Disability severity category**

OPCS weighted severity score	OPCS severity category
0.5 - 2.95	1
3 - 4.95	2
5 - 6.95	3
7 - 8.95	4
9 - 10.95	5
11 - 12.95	6
13 - 14.95	7
15 - 16.95	8
17 - 18.95	9
19 - 21.40	10

Appendix 7

London Handicap Scale (LHS) [1]

This questionnaire is about the way your health affects your everyday life. Please read the instructions for each question and then answer by ticking the box next to the sentence which describes you best. When answering the questions, it may help to think about the things you have done over the last week and compare yourself with someone like you who is in good health.

1 Getting around

Think about how you get from one place to another, using any help, aids, or means of transport that you normally have available. Does your health stop you from getting around?

Not at all - You go everywhere you want to, no matter how far away.

Very slightly - You go most places you want, but not all.

Quite a lot - You get out of the house, but not far away from it.

Very much - You don't go outside, but you can move around from room to room indoors.

Almost completely - You are confined to a single room, but can move around in it

---

[1] Harwood and Ebrahim 1995 (241)

Completely - You are confined to a bed or a chair. You cannot move around at all. There is no-one to move you.

2 Looking after yourself

Think about things like housework, shopping, looking after money, cooking, laundry, getting dressed, washing, shaving and using the toilet. Does your health stop you looking after yourself?

Not at all - You can do everything yourself.

Very slightly - Now and again you need a little help.

Quite a lot - You need help with some tasks (such as heavy housework or shopping), but no more than once a day.

Very much - You can do some things but you need help more than once a day. You can be left alone safely for a few hours.

Almost completely - You need help to be available all the time. You cannot be left alone safely.

Completely - You need help with everything. You need constant attention, day and night.

3 Work and leisure

Think about things like work (paid or not), housework, gardening sports, hobbies, going out with friends, travelling, reading, looking after children,

watching television and going on holiday. Does your health limit your work or leisure activities?

Not at all - You can do everything you want to do.

Very slightly - You can do almost all the things you want to do.

Quite a lot - You find something. to do almost ail the time, but cannot do some things for as long as you would like.

Very much - You are unable to do a lot of things, but can find something to do most of the time.

Almost completely - You are unable to do most things, but can find something to do some of the time.

Completely - You sit all day doing nothing. You cannot keep yourself busy or take part in activities.

4 Getting on with people

Think about family, friends and the people you might meet during a normal day. Does your health stop you getting on with people?

Not at all - You get on well with people, see everyone you want to see, and meet new people.

Very slightly - You get on well with people, but your social life is slightly limited.

Quite a lot - You are fine with people you know well, but you feel uncomfortable with strangers.

Very much - You are fine with. people you know well but you have few friends and little contact with neighbours. Dealing with strangers is very hard.

Almost completely - Apart from the person who looks after you, you see no-one. You have no friends and no visitors.

Completely - You don't get on with anyone, not even people who look after you

5 Awareness of your surroundings

Think about taking in and understanding the world about you, and finding your way around in it. Does your health stop you understanding the world around you?

Not at all - You fully understand the world around you. You see, hear, speak and think clearly, and your memory is good.

Very slightly - You have problems with hearing, speaking, seeing or Your memory, but these do not stop you doing most things.

Quite a lot - You have problems with hearing, speaking, seeing pr our memory which make life difficult a lot of the time. But, you understand what is going on.

Very much - You have great difficulty understanding what is going on.

Almost completely - You are unable to tell where you are or what day it is. You cannot look after yourself at all.

Completely - You are unconscious, completely unaware of anything going on around you.

6 Affording the things you need

Think about whether health problems have led to any extra expenses, or have caused you to earn less than you mould if you mere healthy. Are you able to afford the things you need?

Yes, easily - You can afford everything you need. You have easily enough money to buy modern labour-saving devices, and anything you may need because of ill-health.

Fairly easily - You have just about enough money. It is fairly easy to cope with expenses caused by ill-health.

Just about - You are less well off than other people like you; however, with sacrifices you can get by without help.

Not really - You only have enough money to meet your basic needs. You are dependent on state benefits for any extra expenses you have because of ill-health.

No - You are dependent on state benefits, or money from other people or charities. You cannot afford things you need.

Absolutely not - You have no money at all and no state benefits. You are totally dependent on charity for your most basic needs.

Appendix 8

Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) [1]

This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. Answer every question by marking the answer as indicated. If you are unsure about how to answer a question, please give the best answer you can.

1. In general, would you say your health is (circle one):

Excellent - 1
Very good - 2
Good - 3
Fair - 4
Poor - 5

2. Compared to one year ago, how would you rate your health in general now ?

(circle one)

Much better now than one year ago - 1
Somewhat better now than one year ago - 2
About the same as one year ago - 3
Somewhat worse now than one year ago - 4
Much worse now than one year ago - 5

[1] Ware *et al.* 1993 (56)

3. The following questions are about activities you might do during a typical day. Does <u>your health now limit you</u> in these activities? If so, how much?

(circle one number on each line)

ACTIVITIES	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. **Vigorous activities,** such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. **Moderate activities,** such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
c. Lifting or carrying groceries	1	2	3
d. Climbing **several** flights of stairs	1	2	3
e. Climbing one flight of stairs	1	2	3
f. Bending, kneeling, or stooping	1	2	3
g. Walking **more than a mile**	1	2	3
h. Walking **hall a mile**	1	2	3
i. Walking **one hundred yards**	1	2	3
j. Bathing or dressing yourself	1	2	3

During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

(circle one number on each line)

	YES	NO
a. Cut down on the **amount of time** you spent on work or other activities	1	2
b. **Accomplished less** than you would like	1	2
c. Were limited in the **kind** of work or other activities	1	2
d. Had **difficulty** performing the work or other activities (for example, it took extra effort)	1	2

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional Problems (such as feeling depressed or anxious)?

(circle one number on each line)

	YES	NO
a. Cut down on the **amount of time** you spent on work or other activities	1	2
b. **Accomplished less** than you would like	1	2
c. Didn't do work or other activities as carefully as usual	1	2

)uring the past 4 weeks, to what extent has your physical health or emotional 1 problems interfered with your normal social activities with family, friends, neighbours, or groups (circle one)?

Not at all - 1
Slightly - 2
Moderately - 3
Quite a bit - 4
Extremely - 5

7. How much bodily pain have you had during the past 4 weeks (circle one)?

None - 1
Very mild - 2
Mild - 3
Moderate - 4
Severe - 5
Very severe - 6

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework) (circle one)?

Not at all - 1
A little bit - 2
Moderately - 3
Quite a bit - 4
Extremely - 5

These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks -

(circle one number on each line)

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
a. Did you feel full of life?	1	2	3	4	5	6
b. Have you been a very nervous person?	1	2	3	4	5	6
c. Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
d. Have you felt calm and peaceful?	1	2	3	4	5	6
e. Did you have a lot of energy?	1	2	3	4	5	6
f. Have you felt downhearted and low?	1	2	3	4	5	6
g. Did you feel worn out	1	2	3	4	5	6
h. Have you been a happy person?	1	2	3	4	5	6
i. Did you feel tired?	1	2	3	4	5	6

During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.) ?

(circle one)

All of the time - 1

Most of the time - 2

Some of the time - 3

A little of the time - 4

None of the time - 5

11. How TRUE or FALSE is each of the following statements for you?

(circle one number on each line)

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
a. I seem to get ill a little easier than other people	1	2	3	4	5
b. I am as healthy as anybody I know	1	2	3	4	5
c. I expect my health to get worse	1	2	3	4	5
d. My health is excellent	1	2	3	4	5

Copyright © 1992 New England Medical Center Hospitals, Inc. All rights reserved. (MOS SF-36 English (UK) Version 1.1 10/92)

Appendix 9

General Health Questionnaire (GHQ) [1]

We should like to know if you have had any medical complaints and how your health has been, over the past few weeks. Please answer all the questions on the following pages simply by marking the answer which you think most nearly applies to you. Remember that we want to know about present and recent complaints, not about those you have had in the past. Have you recently:

Item	Response options and scoring			
	score 0		score 1	
Been feeling perfectly well and in good health?	better than usual	same as usual	worse than usual	much worse than usual
Been feeling in need of a good tonic?	not at all	no more than usual	rather more than usual	much more than usual
Been feeling run down and out of sorts?	not at all	no more than usual	rather more than usual	much more than usual
Felt that you are ill?	not at all	no more than usual	rather more than usual	much more than usual
Been getting any pains in your head?	not at all	no more than usual	rather more than usual	much more than usual
Been getting a feeling of tightness or pressure in your head?	not at all	no more than usual	rather more than usual	much more than usual
Been having hot or cold spells?	not at all	no more than usual	rather more than usual	much more than usual

[1] Goldberg 1978 (249)

Item	Response options and scoring			
	score 0		score 1	
Lost much sleep over worry?	not at all	no more than usual	rather more than usual	much more than usual
Had difficulty staying asleep once you were off?	not at all	no more than usual	rather more than usual	much more than usual
Felt constantly under strain?	not at all	no more than usual	rather more than usual	much more than usual
Been getting edgy and bad-tempered?	not at all	no more than usual	rather more than usual	much more than usual
Been getting scared or panicky for no good reason?	not at all	no more than usual	rather more than usual	much more than usual
Found everything getting on top of you?	not at all	no more than usual	rather more than usual	much more than usual
Been feeling nervous and strung-up all the time?	not at all	no more than usual	rather more than usual	much more than usual
Been managing to keep yourself busy and occupied?	more so than usual	same as usual	rather less than usual	much less than usual
Been taking longer over things you do?	quicker than usual	same as usual	longer than usual	much longer than usual
Felt on the whole you were doing things well ?	better than usual	about the same as usual	less well than usual	much less well
Been satisfied with the way you've carried out tasks ?	more satisfied	about the same as usual	less satisfied that usual	much less satisfied

Item	Response options and scoring			
	score 0		score 1	
Felt that you are playing a useful part in things ?	more so than usual	same as usual	less useful than usual	much less useful
Felt capable of making decisions about things ?	more so than usual	same as usual	less so than usual	much less capable
Been able to enjoy your normal day to day activities	more so than usual	same as usual	less so than usual	much less than usual
Been thinking of yourself as a worthless person ?	not at all	no more than usual	rather more than usual	much more than usual
Felt that life is entirely hopeless ?	not at all	no more than usual	rather more than usual	much more than usual
Felt that life isn't worth living ?	not at all	no more than usual	rather more than usual	much more than usual
Though of the possibility that you might make away with yourself ?	definitely not	I don't think so	has crossed my mind	definitely has
Found at times you couldn't do anything because you nerves were too bad?	not at all	no more than usual	rather more than usual	much more than usual
Found yourself wishing you were dead and away from it all ?	not at all	no more than usual	rather more than usual	much more than usual
Found that the idea of taking your own life kept coming into your mind ?	definitely not	I don't think so	has crossed my mind	definitely has

Appendix 10

Mini-Mental State Examination (MMSE) [1]

Record response to each question

Domain tested and test	Score
Orientation	
Year, month, day, date, time	/ 5
Country, town, district, hospital, ward	/ 5
Registration	
Examiner names three objects (for example, apple, table, penny)	
Patient asked to repeat three names-score one for each correct answer	/ 3
Then patient to learn three names (i.e. repeat until correct)	
Attention and Calculation	
Subtract 7 from 100, then repeat from result, etc. Stop after 5.	/ 5
100, 93, 86, 79, 72, 65.	
(Alternative: spell 'world' backwards. D L R 0 W)	
Recall	
Ask for three objects learnt earlier	/ 3

[1] Folstein *et al.* 1975 (251)

Domain tested and test	Score
Language	
Name a pencil and watch	/ 2
Repeat 'No ifs, ands, or buts'	/ 1
Give a three-stage command. Score one for each stage (e.g. 'Place index ringer of right hand on your nose, and then on your left ear.')	/ 3
Ask patient to read and obey a written command on a piece of paper stating: 'Close your eyes'	/ 1
Ask patient to write a sentence. Score if it is sensible and has a subject and a verb	/ 1
Copying Ask patient to copy a pair of intersecting pentagons	/ 1

Appendix 11

Staff-report transition question of change in disability

Name: Study No.

In the opinion of the treating multidisciplinary team, this person has undergone the following change in disability (please tick one):

IMPROVEMENT	a) Minimal
	b) Moderate
	c) Marked
NO CHANGE	
DETERIORATION	a) Minimal
	b) Moderate
	c) Marked

Appendix 12

Effect of the FIM and FIM+FAM training programme on rating proficiency

Objectives

1 To determine FIM and FIM+FAM rating accuracy for individuals and multidisciplinary teams (MDT).

2 To quantify the effect of the formal training programme on the accuracy of individual person rating.

Method

Vignettes (simple, short, written scenarios of patient performance) with model answers were provided by the developers of the FIM+FAM. At the formal FIM and FIM+FAM training day individuals from the three rehabilitation units rated 30 standard vignettes, one for each item of the FIM and FIM+FAM, before and after training. In addition, after training only, clinicians formed MDT's (minimum of three disciplines) and rated by consensus opinion a set of 30 different vignettes. Manuals were provided. Results are reported as percent exact agreement with the model answers and standard deviations (SD).

Results

Sixty one clinicians attended the training days (NRU n = 18, RNRU n = 23, RRU n = 20). Seven clinical disciplines were represented: physiotherapy (n = 16), occupational therapy (n = 11), nursing (n = 18), speech and language therapy (n = 8), neuropsychology (n = 2), social work (n = 3), medicine (n = 3).

Mean percent exact agreement with model answers were:

	Mean percent (SD) exact agreement with model answers	
Type of rating	Pre-training	Post-training
Individuals	61.4% (14.67)	65.9% (13.67)
MDT	-	89% (5.76)

Conclusion

Results indicate that a comprehensive one day training programme did not improve individual rating accuracy for vignettes. However, team consensus rating is far superior to individual rating.