ADL assessment after stroke: aspects on reliability
The stability between raters, instruments and modes of administration

Yvonne Daving

UNIVERSITY OF GOTHENBURG

From the Institute of Neuroscience and Physiology /Rehabilitation Medicine
The Sahlgrenska Academy at the University of Gothenburg
Göteborg 2009, Sweden

ADL assessment after stroke: aspects on reliability. The stability between raters, instruments and modes of administration.

ISBN: 978-91-628-7641-8

yvonne.daving@vgregion.se

From the Institute of Neuroscience and Physiology / Rehabilitation Medicine, the Sahlgrenska Academy at the University of Gothenburg, Göteborg, Sweden

Printed by Geson Hylte Tryck AB, Göteborg, Sweden, 2009

ADL ssessment after stroke: aspects on reliability. The stability between raters, instruments and modes of administration

Yvonne Daving, Institute of Neuroscience and Physiology / Rehabilitation Medicine

ABSTRACT

Activities in daily living, ADL are assessed as an outcome of the rehabilitation process or to be able to determine suitable interventions. The overall aim here was to analyse assessments in terms of ADL in order to investigate influences of ADL assessment and study the stability of the raters, instruments and modes of administration in assessing personal, P-ADL and instrumental, I-ADL items.

Different assessment procedures and instruments were used: the Functional Independence Measure, FIM™, the Instrumental Activity Measure, IAM and the ADL taxonomy. The participants were a convenience sample of stroke victims. The stability in assessing ADL performance was studied according to the following: 1) inter-rater agreement, 2) stability by raters, 3) systematic disagreement on the item level, 4) agreement between different modes of administration and 5) agreement between the ADL instruments on the item level.

Reliability varied for the FIM™ and IAM items, with generally good inter-rater agreement for the same interview situation but less stable agreement in a reproduced semi-structured interview according to the same procedure with a flow chart. There was generally moderate to good agreement between two modes of administration except in some items. It was also possible to maintain a general stability of the assessed ADL dependency after dichotomising the ADL data when a questionnaire was used. Problems related to the instrument and method used (i.e. the assessment procedure) and environmental influences were identified; these were different use of the categories of the scales and interpretation of the concept of ADL independence by subjects´ self-report as conducted by different raters and modes of administration.

The study indicates that further use of a self-reported questionnaire might be an accessible mode of administration in clinical work, both in hospital and in primary care. However, the assessment procedure needs further development to suit each clinical situation, such as acute care. Complementary use of an ADL questionnaire and the semi-structured interview might facilitate clinical interventions, making them more cost effective and reducing the time required for both patients and the professional. This thesis can be seen as a part of further analyses to develop clinically valid and applicable ADL assessment tools.

Key words: outcome assessment, stroke, activities of daily living, reproducibility, questionnaire, interview

ISBN: 978-91-628-7641-8

LIST OF ORIGINAL ARTICLES

This thesis is based on the following four papers, which will be referred to in the text by their Roman numerals:

I. Daving Y, Andrén E, Nordholm L, Grimby G.
Reliability of an interview approach to Functional Independence Measure.
Clin Rehabil 2001 Jun; 15(3):301-310.

II. Daving Y, Andrén E, Grimby G.
Inter-rater agreement using Instrumental Activity Measure.
Scand J Occup Ther 2000; 7:33-38.

III. Grimby G, Andrén E, Daving Y, Wright B:
Dependence and perceived difficulty in daily activities in community-living stroke survivors two years after stroke.
Stroke 1998; 29:1843-1849.

IV. Daving Y, Claesson L, Sunnerhagen KS.
Agreement in activities of daily living performance after stroke in a postal questionnaire and interview of community-living persons.
Acta Neurol Scand, DOI :10. 1111/j.1600-0404.2008.01113x (early view oct 2008)

CONTENTS

ABBREVIATIONS

ICF	International Classification of Functioning, Disability and Health
ADL	Activities of Daily Living
P-ADL	Personal Activities of Daily Living
I-ADL	Instrumental Activities of Daily Living
FIM™	Functional Independence Measure
UDS	Uniform Data System for Medical Rehabilitation
IAM	Instrumental Activity Measure
COPM	Canadian Occupational Performance Measure
NIHSS	National Institutes of Health Stroke Scale
ICC	Intraclass Correlations Coefficient
PA	Percentage Agreement
ROC	Relative Operating Characteristic

INTRODUCTION

One part of clinical research, as in rehabilitation, is to study how outcomes should be assessed to meet the criteria for "overall effectiveness of clinical care" (22) and to identify the individual level of functioning and changes in ability over time (6). It is important in the research to establish the reliability and validity of the assessments under study. The main issue of research, e.g. clinical research, is to have control of the measurement techniques and the eventual assessment variations to ensure that the assessments are reproducible and enable the best conclusions to be drawn (22). Clinical research is a structured process as "it proceeds in a systematic way to examine clinical conditions and outcomes" (71). The data collection procedure in clinical services needs to be audited to ensure that it contains the necessary information about the individual to be able to follow the progress of the disease and identify individual needs (93). At the present time, economic policies and the wishes of consumers have also "obligated the health professionals to define and document outcomes with a focus on end results of patients' care in terms of disability" (71). The use of assessments is also important to helping the individual rehabilitation service to communicate their results to other health care services (93).

During the past decades, there has been extensive international work on the consequences of disease. In 1980, the World Health Organization, WHO presented a classification of the consequences of diseases and injuries, the International Classification of Impairments, Disabilities and Handicaps (95). This was further developed into the International Classification of Functioning, Disability and Health, ICF (96) as a "components of health classification". The term disability is often used according to WHO´s classification of functioning, which is a theoretical foundation for understanding and giving a more universal language when comparing or reporting outcomes (96). The ICF emphasises the positive terms of functioning: namely body function/structure, activity and participation, as well contextual factors including environmental and personal factors (Figure 1). The theoretical approach of the ICF has influenced and inspired the development of concepts in for example assessment instruments and theory models in all the different disciplines of health care. Linguistic norms may facilitate the communication between different professionals at different health care services, e.g. acute medical care and primary health care in the community.

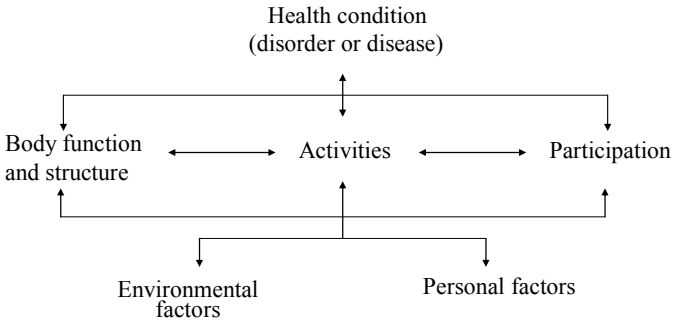

Figure 1. The model of International Classification of Functioning, Disability and Health, ICF, "the biopsychosocial model", WHO 2001.

While medical services have a great focus on the consequences of disease in terms of impairment and activity limitations, the public health services in the community are usually more involved in activity, participation and the context, e.g. environmental factors. In the different environments, the methods used usually reflect a more medical or a mainly social-psychological approach, including cultural norms (58, 88).

ICF is based on an integration of the two models, the medical model and the social model, into a "biopsychosocial" model, viewing "various perspectives of functioning" as a "complex collection of conditions" directly caused by disease or as a problem created by the social environment (Figure 1) (96).

The ICF has become a well-known tool to connect and classify the assessment instruments used in health care. For example, the descriptions of the items in the instruments can then be interpreted and analysed and misfitting items can be identified and removed to create a uni-dimensional scale. In this research process it is important to map different constructs and domains to create models, to study different aspects of the process of developing new instruments or to identify core sets for treatment programs (96). The definition of activity "is the execution of a task or action by an individual, it represents the individual perspective of functioning" (96). The components of the ICF model describe an individual's functioning in a specific domain as "a complex relationship between the health condition and contextual factors"(i.e. the physical, social and attitudinal environments). According to the ICF, the environmental components and the health condition are among the factors that influence activities in daily life (Figure 1). Capacity "is a construct that indicates the highest probable level of functioning that a person may reach in a domain at a given moment" or is understood in terms of "executing tasks in a standard environment" (without adaptations). Capacity is measured in a "uniform or standard" environment (96). The performance component "describes what an individual does in his/her current environment". Because the current environment includes a societal context, performance can also be understood as "involvement in a life situation". This

includes all aspects of environmental factors (physical, social and attitudinal) (96). The ICF classifications are widely used to analyse instruments and treatment programmes in rehabilitation units.

One basic issue is to measure and evaluate outcomes in rehabilitation (6), e.g. functional assessments, which are fundamental in clinical practice and research (88). Activities of daily living, ADL contain activities that are usually performed in everyday living and are necessary for community living and assessments of ADL are often occurring in rehabilitation. Outcome studies in occupational therapy are also being carried out and the ADL instruments and taxonomies used are under critical review (56, 58). Conceptual clarity is important for identifying changes between two measurements, e.g. responsiveness, to form the basis of an evaluation of the assessment (82, 92) and consider and choose the most suitable instrument for actually assessing the particular phenomenon in question (93). The evaluation of disability also depends upon the diagnosis (6, 69). There is a need to understand the interaction between the patient and the environment and whether it is possible to obtain optimal outcomes under defined clinical conditions and to find methods of observing human behaviour in a reliable and valid way (71).

Concepts of measurement

Methodological research aims to document and improve the reliability and validity of the instrument used in measurements or assessments and the assessment procedure employed (22). Analysing features of ordered categorical data often, in contrast to numerical data, requires non parametric statistics (84). The numerical data refer to the accuracy and precision of the measurements; corresponding expressions for categorical data are reliability and validity. Regarding assessments of ordinal data, reliability refers mainly to the quality of the raters using the instrument and validity refers to the quality of the instrument (84). The guidelines are the most important concept in the measurement procedure because they determine the quality of the measurement (65).

In both clinical and research settings, it is desirable to use reliable and valid ways to monitor the assessments used. The assessments are needed to describe and predict the ability of an individual at a specific moment, to evaluate changes over time and to examine the effectiveness of interventions (58). Variability, as a source of error in an assessment, may arise from the rater (i.e. lack of consistency in the individual rater, intra-rater), the measuring process or the raters (lack of consistency between raters, inter-rater) (7). The latter author emphasised that the variations might be inherent in the method itself. This is a problem in both clinical and research settings, as assessments are made by different raters using the same method.

The presence of intra-rater disagreement, bias, in a repeated assessment indicates systematic differences, and this disagreement might arise when there are different interpretations of the

definitions of the categories in the instrument or different self-reports by a respondent on two different occasions (17, 84). Random and systematic differences between raters may occur in any situation that includes observations (14, 17, 33, 87). The rater's interpretation of the guidelines, as well as the personality, prejudice and the experiences of the rater, can give rise to systematic differences (84).

The structure of measurement

The nature of measurement can be viewed as a procedure in which numbers or letters are assigned to properties according to special rules, guidelines (65). The numbers are the results of the measurement and are useful for comparisons and evaluations. In the measurement procedure of the ADL assessment, the observed disability or the disability assessed in an interview is transformed into ordered categories, often defined in a guideline. Many categorical items have an ordered structure and a distinct order, but the distance between the categories is not known (84). These categories in most ADL instruments or taxonomies are ordered categories, and give *ordinal data*, with some kind of relation. Typical relations are more or less independence/dependence, e.g. assistance from another person (60) or assistive devices (67). Another typical relation can be perceived difficulty (38). In ADL assessments using ordinal data, the therapist makes an attempt to categorize abilities according to special rules or guidelines for the instrument (82). The ADL checklists, which do not have structured ordering, between the items, give *nominal data*, for example when dichotomising the data and calculating frequencies. The *interval* level of measurement has, besides hierarchic ordering; also an equal distance between observations, e.g. one can specify how many units greater one observation is than another. Rasch analyses transform ordinal data to linear interval data.

Validity

Validity can be defined as the extent to which an instrument measures what it is intended to measure (65) or "the meaningfulness of test scores as they are used for specific purposes" (22). The problem of validity in measuring health variables has to do with the estimates being in general of indirect behaviour, and the rater/observer is never completely certain that he or she is assessing the precise property (65).

There are different types of validity that can be investigated to see whether the assessments are valid, that is, how near the true state the assessment is. For persons living outside institutional settings, it is important that the assessments including areas of interest for independent living in the community. The construct validity is "the validity of the abstract construct that underlies measures, not directly observable" (22). The construct of the ADL instrument is important for the validity, and the items included in the instrument should cover essential areas of the construct. In developing

instruments, the initial stage is to explore and select relevant activities/items and to decide how to define them. One way to examine the construct validity can be to use the Rasch model in order to deal with the instrumental structures that give rise to disordered thresholds (66, 98).

Reliability

Reliability is not a uniform concept, and it can be estimated in several ways: 1) test-retest (intra-rater), 2) inter-rater reliability, 3) stability of the measure on repeated administration and 4) internal consistency (22, 65, 70, 71). A strict definition of reliability is "the extent to which a measure is free from error" or, in other words, "the extent to which measurements are repeatable" (1, 22) or "the degree of consistency with which an instrument or rater measures a variable" (71).

Categorical data, the types of data used in for example the ADL instrument, are not standardised and are suitable for non-parametric statistics. Reliability is by definition one part of the analysis of relationships, including different measures of agreement and correlation (22, 71). Rater or method agreement is used to study whether one observer or method can be replaced by another without changing the outcome, thus making it possible to investigate the reliability of the results. The agreement between the results, such as found by kappa statistics, can been used to analyse the consistency or stability between raters and methods (84).

The components of reliability include the instrument/method, test situation, the rater and the intra-subject variability (22, 84). The usefulness of an instrument depends highly on the degree of reproducibility of the assessment as used by different individuals. For this reason, it is important to question whether the assessment generates consistent, reproducible information (1). Studies using more structured procedures and/or methods are assumed to be more reliable (22).

Repeatability expresses the minimum variability of a test-retest (within the same rater) and reproducibility expresses a maximum observer (between two or more raters) variability (22, 84) in inter-observer assessments.

The rater´s or participant's (e.g. in self-reported methods) use of each category can increase both the intra-rater´s and the intra-participant's disagreement in instruments with several ordered categories and, here, systematic disagreements (bias) will occur (84). With different raters there is also a possibility that systematic disagreement will occur when a rating scale is used (7). A non-parametric statistical method has been developed for paired, ordered data to analyse systematic disagreements and occasional disagreement, and is used in the analysis of aspects influencing self-reported ADL assessments (17 21, 84).

Assessment of persons after stroke

One aim of rehabilitation efforts is to help the individual after stroke to attain optimal physical and cognitive ability so that he or she is able to return to the own home, outside institutions. Stroke

strikes more than 30 000 persons every year in Sweden and is the most common cause of new disabilities in the adult population (80).

The clinical symptoms after a stroke might be complex, involving several functional neurological systems (3). The localisation of the brain damage gives the observer information about how suitable assessments should be initiated for both physical and/or neuropsychological consequences. A lesion in the right hemisphere is often associated with impairments in visuo-spatial perception, and a lesion in the left hemisphere (if dominant) might lead to linguistic problems. Tiredness, concentration and attention deficits and other cognitive impairments may interfere with occupational performance in activities of daily living. One commonly reported consequence of stroke is fatigue (9), which seems to be associated with greater dependence in personal care activities, P-ADL (9, 32, 94).

Impaired cognitive function can lead to difficulties in the interview situation in interpreting responses. For example, aphasia and memory problems can mean that there is inadequate information or misinterpretations of problems, which demonstrates the importance of sufficient knowledge of the medical condition. There might also be a problem with overestimating or underestimating the ability to perform ADL. The advantage of the interview approach is the possibility it gives the rater to add complementary questions related to the patient's behaviour or to directly observe the patient's reactions (71). After a stroke, a person might be helped by the interviewer concretising the situations in which the person performs daily activities with examples or questions. The patient's inconsistent performance as a result of the stroke is difficult to control and distinguish from others factors, leading to rating variability (22). The ability is supposed to lie in the person himself or herself, as a consequence of stroke impairment or in something related to other intra-personal or contextual factors (88). Assessments of ADL do not usually give attention to diagnosing specific symptoms.

It is important to describe an individual's ability to perform daily activities after a stroke and to be able to discriminate changes in the progress, and this is needed in interventions. After a disabling event such as stroke, the ability to perform personal care must often be assessed and trained before complicated and more demanding instrumental activities can be performed.

The ADL instrument

The aim of all rehabilitation services must be for patients to reclaim the best possible health and ability as soon as possible. The methods used, including an effective assessment procedure, must be continuously reviewed and modified to suit busy clinical settings. The instruments chosen are often used for discriminative, predictive or evaluative purposes (53, 67, 82).

Historically, it has been a challenge to measure ADL performance. Neither is there any consensus as to what method should best be used to describe ADL on either on an individual or a group level (20, 50). One problem in assessing ADL performance is the normal fluctuation and a volatility in daily

performance that Smith called “a moving target phenomenon” (79) or the instability of the nature of ADL ability (58). The latter author stated that there is a need for future research to examine the clinical utility of ADL assessments to be able to know how an instrument can be used in different settings and by different raters.

Since 1935, when the first ADL checklist was presented to identify common activities according to safety and independence, many attempts have been made to study daily activities. Different ADL checklists or ADL instruments were developed to assess personal care activities, P-ADL and instrumental activities, I-ADL.

Two early ADL instruments that assess independence were developed to comprise personal care activities, the Barthel ADL index (63) and the Katz ADL index (51), in order to assess ADL independence. The Barthel index uses a two to four-step level to rate dependence in P-ADL. The ADL staircase assesses dependency with four instrumental activities and was developed as a supplement to the Katz ADL index with six personal care activities (45). These two scales are built on different concepts; the Barthel index has a more empirical perspective and the Katz index is built on theories of human development and activities are studied hierarchically to form a Guttmann analysis (51, 63)

The ADL concept was later extended to consider activities necessary to maintain living in the community, e.g. instrumental activities (59). I-ADL´s are expected to be more susceptible to the influences of environmental factors (e.g. roles), as well personal interests and motivation, than are the personal ADL items (82).

Many recent ADL checklists favour performance, what the person actually does, compared with the original Barthel index, which considered what the person “is able to do” (64). This approach was changed during the 1980s to form a modified Barthel index, “what does the person do” (16). Asking what the respondent “can do” may provide a hypothetical answer that records what a person thinks he can do. This approach can exaggerate the ability of the respondent by as much as perhaps 15 - 20% (64).

The ADL instrument often assesses some aspect of performance. Besides the assessment of independent performance in personal and instrumental activities (37, 46, 81) different aspects of occupational performance might be added, such as perceived difficulty (36) and safety aspects (17, 69). It is also possible to add different aspects such as satisfaction (2), use of assistive devices and/or altered working methods (67). In clinical OT practice it is important to use, besides ADL descriptions of personal care, client-centred ADL assessments, such as the Canadian Occupational Performance Measure, COPM to improve decisions in specific patients or treatment programs (57). COPM is an assessment intended to identify the individual priorities of activities of daily living and how important they are in relation to one another. In the assessment of the individual ability to perform ADL it is important to have a client-centred perspective (97) to distinguish what is a chosen

ADL dependency that does not need any intervention from problem areas that are in need of rehabilitation interventions (5).

In the late 1980s the Functional Independence Measure, FIM™ was developed by a task force from different rehabilitation organisations (40) with the aim to create a broader and more sensitive instrument than had earlier been available. The FIM™ was intended to be a measure of the burden of care, discriminating verbal assistance from physical assistance in a seven-step scale. The FIM™ instrument then showed two separate indicators of disability, a motor and a social-cognitive part (60). Further studies of the construct validity on the FIM™´s original seven-step scale showed it to have disordered categories in both physical (37) and social-cognitive items unless reduced according to Rasch analysis (61).

In developing an applicable instrument for use with community living persons, items of the FIM™ were combined with instrumental items generating the Instrumental Activity Measure, IAM (37, 38). These additional instrumental items were intended to eliminate the ceiling effect of the personal care instruments on the outcome, to make the measure more sensitive and discriminate the individual's ADL ability to continue to live outside an institution. IAM assesses the patient's ADL ability to perform selected activities that are common and necessary to live in the community. It was developed to be used in interview form and is added in parallel with the subject's perceived difficulty during the interview. The inter-rater agreement was shown to be good, but the validity needs further investigation, as does reliability when used by different raters and in different settings (19).

The ADL taxonomy was developed by occupational therapists, OT, for OTs in clinical situations as a guide for observations and/or in interviews (82). One of the efforts of OTs´ was to develop the concept of ADL to operationalize and categorize activities of daily living useful for clinical observations. The ADL taxonomy is a "systematic classification" of common activities to assess ability in different activities in order to describe overall performance in daily life (89). The taxonomy has been analysed with regard to its content validity, and the ordinal properties have been studied, resulting in ordered actions (parts of activities) from easy to difficult actions (82).

The ADL assessment

ADL assessments can be used to measure individual outcome at the hospital or after a rehabilitation programme or can be collected as one part of other health outcomes for longitudinal studies, e.g. databases. Predictions of the eventual need of assistance in daily activities are important for the planning of the discharge from hospital. Historically the first outcome instruments measured the more basic P-ADL. A more complex set of items is now included in different ADL assessments, such as patient satisfaction or participation. This complexity of different items and aspects of

assessments creates difficulties in comparing the results between different settings, and several instruments (included ADL) have been shown to be setting and situation specific (50).

Occupational performance and assessment of activities of daily living are core elements of occupational therapy. The definition of "occupation" is a goal-directed use of time, energy, interest and active participation in ADL, work and leisure, e.g. performance areas (69). ADL ability is the "ability for occupational performance; to perform (to do)" (89). It is important to identify self-perceived health as it influences occupational performance after stroke (58). ADL assessments have been used among other variables to document outcome, i.e. the result of rehabilitation efforts.

ADL includes activities that are usually performed in everyday life and are necessary for community living. P-ADL include activities such as dressing, feeding and transfers. Activities demanding more complex abilities are the I-ADL, such as cooking, shopping and transportations. The instrumental activities require more advanced problem-solving skills and are also influenced by social skills and habits (69). The ability to perform activities thus differs with both individual capacity and setting. However, all activities are to some extent individual and interfere with environmental demands (socio-attitudinal and physical demands) in hospital or in the home context. The environment might demand a more rapid or qualitative ADL performance, resulting in the need of assistance from another person, where the physical obstacles in the environment are the predominant problem.

There is no uniform definition of the ADL concept and what activities the assessments should comprise for them to be valid; it depends on the aim (58, 82). Neither is there agreement as to which activities the ADL assessments should contain to be valid for clinical or research use (56, 58, 89). There are variations in its use and there is still debate about what activities should be included in the ADL assessment and what the most important areas for examination of ADL are (50, 58).

The assessed concept: Independence

A common performance approach in clinical practice is the "doing" of activities, and the assessment often uses as its concept the "independency or dependency from another person to perform daily activities". It is fundamental to assess the individual ability to perform daily activities, such as in/dependency in P-ADL and some of the I-ADL, at the hospital. The measurement procedure is the same regardless of whether the rater assesses a directly observable variable or must assess indirect, more latent variables such as ADL independence (65, 88). Christiansen pointed out a continuum between "direct observable and measurable behaviour" such as grip strength and postural control and "indirectly observed behaviour" for example memory and interests (13). A continuum of different behaviour has also been explained by Tesio as an "evident/totally present variable" in the "latent/deduced behaviour" such as self-sufficiency in activities such as walking and dressing (88). The latter is "latent and in large part unpredictable and can underline various behaviour at different

moments" and the ADL independence is expected to be such a latent trait of ADL ability (88). In this case the author stated that the assessment is an estimate and cannot be an exact measurement. It follows that it is more difficult to formulate guidelines for the assessment, since the rater must rely on his or her own experience in the interpretation of behaviours, i.e. subjective assessments (64, 65).

Factors influencing the ADL assessment

The factors that influence the assessment procedure of any measurement, including assessment of ADL in a subject, are multifaceted (Figure 2). The factors depend on several sources related to subject, rater, method and test situation (53, 54, 84). All activities of daily living are not pertinent to all persons and this absence of a standardized concept, both in terms of the content of items and ADL performance, makes the ADL assessment more complicated for the rater (58).

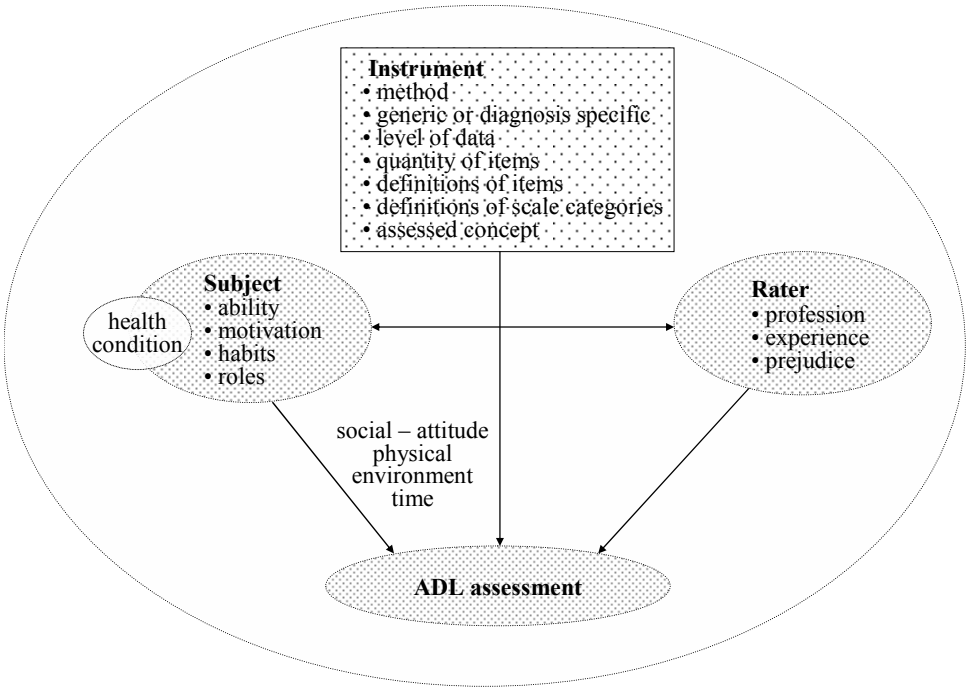

Figure 2. Illustrations of possible factors influencing the ADL assessment.

Difficulties are experienced in interpreting and grading behaviour because each rater must rely on his or her own conclusions in interpreting what step category in the guidelines is most suitable. The complexity might increases when several categories are used by raters in different settings to assess health status. These facts can lead to disagreement between raters. Training in how to use the instrument has been shown to increase within-professional agreement of the assessments with the FIM™ when done by occupational therapists (30) and training is also followed by trained clinicians at medical rehabilitation wards (39). Raters' prejudice about for example the mode of administration also influences the assessments (84).

The consequences of a neurological condition such as stroke might also have a multifaceted individual impact on ADL ability depending on the effects on body function, activity level and the possibility to participate in daily life situations. These above mentioned confounding factors in estimating individual functioning may result in difficulties achieving stable base-line assessments of ADL, where there can be a rater related variability that affects the reliability of the assessment (84). Furthermore, there are several reasons for the complexity and the inconsistency of daily activities, which include the broad range of disabling consequences of stroke (94), individual preferences, choices, environmental demands and socio-cultural habits and routines. It is unlikely that it is possible to be in control of all these inconsistent factors. However, the purpose of an outcome assessment must be to produce reproducible and reliable data to follow real changes in the progress of a health condition and should not be dependent on different sources of errors in the assessment procedure (22).

There are few studies in the literature that analyse the stability of different ADL items and compare different modes of administration of testing ADL ability as concerns in/dependent living after stroke. There are a number of different ADL instruments/checklists/taxonomies in use, which makes it difficult to find and choose the most clinically useful, reliable and valid method that is easily accessible and uses a cost-effective assessment procedure (method and instrument). All assessments that use instruments to structure the assessment procedure are formalized, and guidelines help to make a focused categorisation of the behaviour being assessed. Problems emerge when the rater must identify the latent or hidden behaviour of a subject according to the guidelines.

Environmental influences may be more complicated in the interplay with other persons, e.g. when two or more persons who live together share many of the activities. This might influence the actual assessment as well the evaluation of changes over time (82). There is also an individual time and situational factor that changes unpredictably depending on the socio-demographic situation (82). The environmental and socio-cultural influences on the ADL assessment may be confounding factors that are difficult to control in the "performance approach", as was noted in a study of community living persons after stroke (38). The instrumental activities might especially not be sensitive to change depending on the variability of influencing factors (82).

It is important to analyse ADL assessments in order to study their stability and to gain a deeper knowledge of the data administration procedure after stroke. The challenge for future research is to find an ADL assessment procedure that meets the needs and criteria to be used by different raters in different clinical and research settings (58).

AIMS OF THE STUDY

The overall aim was to analyse self-reported ADL ability in order to investigate the stability of the raters, instruments and modes of data collection in assessments of P-ADL and I-ADL items.

The specific aims were:

- to analyse the consistency of paired semi-structured interview assessments of FIM™ items between two pairs of raters (the same interview conducted with one week between interviews) according to in/dependence in ADL performance
- to analyse the consistency of a paired semi-structured interview assessment of the items of the IAM between one pair of raters during the same interview according to in/dependence in ADL-performance
- to analyse systematic differences between raters in items from the FIM™ and IAM instruments
- to investigate the structure, dimensionality and changes in the hierarchical order of items of the FIM™ and IAM made at discharge and approximately two years later in the home as a part of a follow-up study
- to compare two modes of self-reported in/dependent ADL performance (postal questionnaires and a semi-structured interview) according to the FIM™, IAM and ADL taxonomy
- to compare the results in order to discriminate the person's ADL in/dependency as assessed by each of the ADL assessment tools (FIM™, IAM and ADL taxonomy).

METHODS

Participants

Eighty-one persons, consecutively recruited over a two-year period, were included in the follow-up studies. The persons had undergone rehabilitation for stroke at the Department of Rehabilitation Medicine, Sahlgrenska University Hospital, and had been discharged to their homes. They were invited to participate in the follow-up studies two years (Studies I-III) and 11 years (Study IV), respectively, after stroke onset. Sixty-eight persons (44 men and 24 women) participated at ~ two years and 36 persons (22 men and 14 women) at ~ 11 years after stroke onset. The reasons for not taking part were (drop-outs from the invited group, n=80): one person had moved and two persons did not answer letters or telephone calls. Three persons could not participate because of a new diagnosis and one person was living in a nursing home. Two persons had severe aphasic problems. Three persons declined follow-up at two years after stroke (Figure 3).

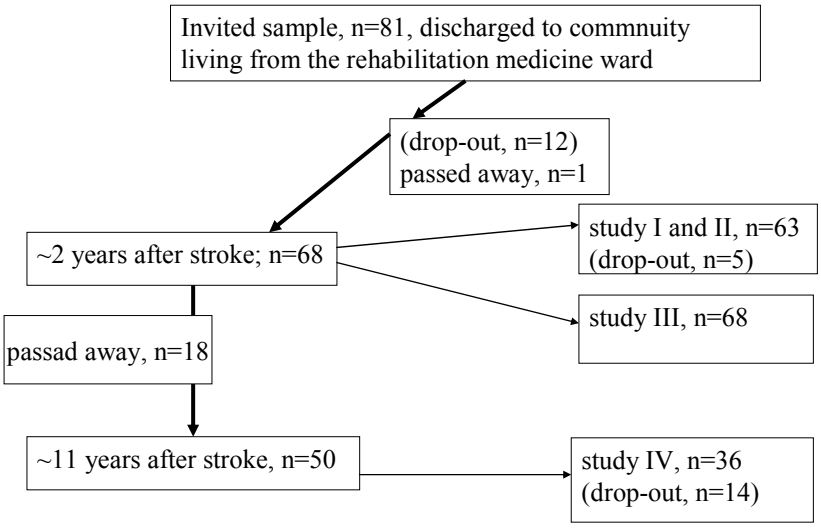

Figure 3. The participants in the studies.

The median time from the onset of stroke to admission to the rehabilitation ward was 30 days (mean 46, SD 46 days). The length of the hospital stay ranged between eight and 210 days (mean days 74, SD 44, median 62). The length of stay did not differ between patients aged ≥55 and <55 years. Sixty-one percent had cerebral infarction, 14 % intra-cerebral haemorrhage and 25 % had subarachnoidal haemorrhage. At discharge, 32 % had right hemiplegia, 57 % left hemiplegia and the remaining patients either a bilateral paresis or no remaining paresis. The neurological deficits according to the National Institutes of Health Stroke Scale, NIHSS were collected by a physician during the visit at the two-year follow-up at the outpatient rehabilitation medicine clinic (Table 1).
Eleven years after stroke onset, the follow-up group at two years was invited to participate in a follow-up study analysing ADL ability. When the persons who had passed away were removed from the group, 36 persons fulfilled the criteria for being able to maintain living in the community, outside institutions (Table 1).

Table 1. Neurological deficit described according to NIHSS. A low score indicates a low impairment score; the maximum score is 34 points.

	Studies I-II (n=63)	Study III (n=68)	Study IV (n=36)	Study IV (drop-out n=14)	Study IV (passed away n=17)
mean	5	5	4	5	6
SD	3	3	3	3	4
median	4	4	4	4	6
range	0-15	0-15	0-11	1-9	0-15

To ease comparisons with other study groups described in the literature, sum scores for the physical and social-cognitive items, respectively, of the FIM™ were used at approximately two years (Table 2) and 11 years (Table 3). These two subscales were separated according to their different constructs (60). They indicate a rather low degree of dependence in physical and social-cognitive items.

Table 2. FIM™ sum score at ~2 years after stroke (n=68) for physical and social- cognitive items respectively. The number of total sum score for the physical items of the FIM™ is 91 and the social- and cognitive items are 35 points.

	Physical (91 p)	Social/cognitive 35 p)
mean	76	25
SD	±13	±6
median	80	26
range	20-91	9-35

All but one of the participants lived in the community. This person lived in the community two years after the stroke but at an institution at 11 years after stroke (Study IV).

Table 3. FIM™ sum score at ~11 (n=36) years after stroke for physical and social-cognitive items respectively. The total sum score for the physical items is 91 points and for the social-cognitive items 35 points.

	Motor (91p)	Social/cognitive (35p)
mean	81	27
SD	±7	±5
median	83	27
range	60-91p	14-35p

Participants in Studies I-III

Sixty-eight persons (44 men, 24 women) participated in the follow-up study two years after stroke; 27 persons had a left hemisphere lesion, 29 had a right hemisphere lesion, 11 persons had bilateral lesions and/or lesions in the basal ganglia/brainstem /cerebellum and one person had both a left hemisphere and a brainstem lesion (Study III, n=68). Due to incomplete data in five persons, the data collected in 63 persons were used in Studies I-II, including 26 persons with a left hemisphere lesion, 26 with a right hemisphere lesion and 11 persons with bilateral and/or basal ganglia/brainstem lesion/cerebellum lesions (Studies I-II, n=63).

Participants in Study IV

Fifty persons from the original group of 68 persons (of whom 18 persons, 13 men and five women, were deceased and another 13 persons declined to take part) were invited to take part in Study IV, which took place approximately 11 years (mean 11 years, 10 months) after stroke onset. A total of 37 persons (22 men and 15 women; mean age 62 years, SD 8) participated in Study IV. However

the data of one person (a woman) were excluded since she was living in a nursing home because of severe cognitive disability and ADL dependence. ADL assessments made in 36 persons were thus used (Figure 4). Ten persons had remaining communication difficulties at the time of the study. Approximately one-fifth had totally restored two-hand function, and the same proportion had no function in the impaired arm/hand. Approximately one-third were living in a single-person household. Six persons had returned to paid work. Cardio-vascular problems were seen in one-third of the participants; only five had musculo-skeletal problems and two persons had confirmed psychiatric problems.

Instruments

The FIM™ (40) and instrumental activities IAM were used to assess the participant's ability to perform activities necessary for independent living in the community (36). The last study (Study IV) used the ADL taxonomy, which was developed by occupational therapists. In a theoretical framework the ADL concept was operationalized to a taxonomy for use by OTs in clinical situations as an observation or interview guide (89). An overview of the assessment tools is shown in Table 5.

Table 4. Assessment tools used in the studies

Assessment tools	Study
FIM™	I, III, IV
IAM	II, III, IV
FIM™/IAM	I, III, IV
FIM™/IAM, ADL taxonomy	IV

The Functional Independence Measure, FIM™

The FIM™ instrument is a generic, internationally used instrument. It was devised by the American Congress for Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation to be used as an assessment tool in the Uniform Data System for Medical Rehabilitation, UDS. The FIM™ was originally developed for observations commonly used in inpatient rehabilitation but is also recommended as a follow-up instrument (36, 68). Its reliability and validity has been documented in several studies since it was introduced in the late 1980s (40, 68). Its internal consistency (the items' homogeneity to assess the characteristic to be studied) has been found to be high (21, 83). Its construct has been studied and has given two different indicators of disability: 13 motor, also called physical, items and five social and cognitive items (41, 60, 83). It consists of items assessing self-care, sphincter management, transfer, locomotion, communication, social interaction and cognition.

The FIM™ activities assess an individual's need of assistance in performing daily activities. The measurement procedure uses a seven-step scale anchored by the extreme ratings of total dependence as "category 1" and complete independence as "category 7". There are two independent categories: "complete independence" (category 7) and "modified independence" (category 6), the latter being assessed when the person needs assistive devises, uses more than "reasonable time" or there is a concern for safety". Further, there are five dependent categories: "supervision" (category 5), "minimal contact assistance" (category 4), "moderate assistance" (category 3), "maximal assistance" (category 2) and "total assistance" (category 1), (Figure 4).

The FIM™ is a measure of disability and is intended to measure what the person actually does and the need of assistance in each item. It was originally developed as a data core set for rehabilitation medicine to assess the burden of care for use at medical wards. The most common use of the instrument is in observations of the subjects (40), but FIM has also been used for interviews (11, 47, 78) and self-ratings in a self-reported questionnaire (35). The FIM™ manual includes a format for semi-structured interviews for each item with a "decision tree" to be used at follow-ups by telephone or to guide clinical observations to determine the most suitable assistance level (91). The FIM™ was designed to be discipline-free, "that is, a measure usable by any trained clinician regardless of discipline". It is intended to measure what the person actually does in common activities in daily life. The lowest scores (e.g. more dependent score) should be chosen if the rater is doubtful of a suitable category level. The collection of the interview data followed the ordinal seven-step scale. However, modified categories were used in the questionnaire form, consisting of five levels.

The Instrumental Activity Measure, IAM

IAM was introduced in 1996 as a supplement to the items of the FIM™. Seven instrumental activities of daily living, I-ADL were arranged and analysed (36). Acceptable inter-rater agreement was shown in a study that used paired independent assessments of I-ADL during the same interview (19). However the authors stated "if the assessments are to be used to evaluate the treatment planning process it is essential to increase the kappa values to above 0.75 to identify individual deficits and assets". The validity is still only illustrated to a limited extent; however, Study III indicated two clusters of item difficulty. Three items, *Locomotion outdoors*, *Simple meal* and *Small-scale shopping,* were clustered separately from the other five items as it was easier for this sample to achieve higher (independent) categories. The other five items were clustered in the same way and were ranked "harder" or activities in which it was more difficult to be independent. There were further some gender differences in the items' difficulty. For example, men found it more difficult to be independent in such activities as *cooking* and *cleaning* than women, while the opposite was true for *Small-scale shopping* and *Locomotion outdoors* (38). The instrument was further analysed in its

instrumental structure according to the hierarchical order of the items in persons after stroke, resulting in a division of the *shopping* item into two items: "*Small-scale shopping*" and "*Large-scale shopping*" (38). The structure of the items follows a similar design and form as the FIM™ with the ordinal seven-step scale for performance and need of assistance. Besides the need of assistance, the person also rates his/her perceived difficulty in the performance of the activities using a self-report ranked in four categories. The actual manual version of the instrument consists of eight activity items that assess commonly selected activities performed to maintain independent living in the community: locomotion outdoors, simple meal, cooking, public transportation, small-scale shopping, large-scale shopping, cleaning and washing (Swedish version 2.1 2003). The assessment uses a semi-structured interview approach that helps the respondent to explain how the eight instrumental activities have been performed during the last month, i.e. with or without assistance. Its measurement properties are ordinal and the usual non-metric statistics for ranked ordinal data should be used.

The ADL taxonomy

The ADL taxonomy was introduced in 1994 as a classification system of activities of daily living (89) and was further developed and analysed according to the operational definitions established in 1999 (82). The theoretical framework operationalized the ADL concept to the ADL taxonomy for use by OTs in clinical situations as a guide for observations and/or interviews (89), and, as the authors pointed out, "it was the first step in the process to develop an assessment instrument based on a consistent concept of ADL". The ADL taxonomy has tested construct (89) and, in addition, a new study has confirmed its content validity (82). Thus an ordinal structure was investigated. In the latter study, different sub-groups of diagnoses of patients for use by OTs in clinical situations were analysed and patients were ranked from the most able to the most disabled (82). The results showed that the ordered structure within the activities indicated a good stability of its construct in the studied diagnosis groups. It is expected to be used in clinical practice on both an individual level and on a group level.

The ADL taxonomy contains 12 common activities in self-care, home maintenance and communication. The taxonomy comprises 12 activities: eating/drinking, mobility, going to the toilet, dressing, personal hygiene, grooming communication, cooking, transportation, shopping, cleaning and washing (90). It uses hierarchical descriptions of ADL performance in daily activities with a central superior concept for each activity. Each activity consists of two to six actions in a rank-ordered structure. Each activity comprises ordered actions, categorised with a label depending on the ability to perform the whole described activity. The recording in each action was dichotomous and labelled (+) for ability to perform (actually do) the actions and (-) for disability (actually do not do) the actions.

An example of the ordered structure of the action "going to the toilet" is presented in Table VII. The activity comprises four actions: 1) *Bowel and urine elimination volitional*, 2) *Getting on and off the toilet and managing oneself after elimination*, 3) *Arranging clothes and equipment such as pads and sanitary towels*, *and washing hands*, and 4) *Getting to and from the toilet in time*. Category A is used when all actions are performed, category B if the most difficult action is not performed and so on (Table 7). In the present study "Ö" is used when, according to the manual, when the rank ordering of actions is disrupted, while, despite this, the persons are assessed as dependent. Additional information about the operational definitions and procedure has been presented (82). The ADL taxonomy does not include social interaction, problem-solving or memory items. Study IV used the 1999 manual version III (90).

Table 5. The ordered structure in the. "*Going to the toilet*" activity, comprising four actions.
Recording in each action was dichotomous (binary) and labelled (+) for ability to perform (actually do) the actions and (-) for disability to perform (actually do not do) the actions.

Actions				
1	2	3	4	Categories
+	+	+	+	A
+	+	+	-	B
+	+	-	-	C
+	-	-	-	D

The questionnaires in the ADL assessment tools

Questionnaires with items from each of the ADL assessment tools were created according to the operational definitions of the items. Items from the FIM™ combined with instrumental items of IAM and the ADL taxonomy were used. Two items from the FIM™ were divided to make them easier to fill in from the subject's perspective: "*Transfer shower/bath*" and "*Walk/wheelchair*".

Questionnaire with items from the FIM™/IAM

In the questionnaire, which was designed with item definitions from the FIM™ and IAM manuals, we used five categories (two independent, "independence with and without assistive devices", and three dependent). The two items from the FIM™, "*Transfer to shower/bath*" and *"Walk/wheelchair",* were divided into two items/questions to make them clearer to the participants. The questionnaire contained 15 physical activities/items (instead of the original 13 physical items) and five social and cognitive items of the items from the FIM™ complemented with the eight activities from the IAM. The questionnaire drawing from the FIM™ /IAM contained 28 items

Questionnaire with items from the ADL taxonomy

The questionnaire version followed the ADL taxonomy but added instructions about how to complete it. The information included and emphasised what independent performance is: "without assistance from another person". All other performance of the activities was interpreted as "dependent on another person to perform". It contained 12 activities/items that included 47 different parts of activities/actions.

Data collection and assessment procedure

The general procedure in the studies was as follows: the participants were contacted by mail and then by telephone to give information about the aim and procedure in their participation.

A semi-structured interview procedure was used, "with latitude for the interviewer to clarify questions as needed for the participants thereby obtaining more information" (22). In each interview, the participant described the performance of each activity according to activity definitions in the instruments to give sufficient information for the interviewer to be able to identify the most suitable "in/dependency" category. If the rater was unsure of a suitable category, the category indicating higher dependence was to be chosen in concordance with the guidelines of the FIM™. In the studies with paired assessments, one of the raters conducted the interview while the other rater listened/or added some clarifying question. The ADL interview approach for the rater was structured with the help of a flow chart ("decision tree") to assess each item. This flow chart is described in the guidelines of the FIM™ (Figure 4) and was used in all persons in the studies except those involved in the interviews concerning the ADL taxonomy. The ADL taxonomy uses a nominal scale to assess a suitable category for the ability to perform activities of daily life. A modified flow chart to suit the instrumental activities of the IAM is used as given in the guidelines for the IAM. The same flow chart also guided observations in the hospital setting.

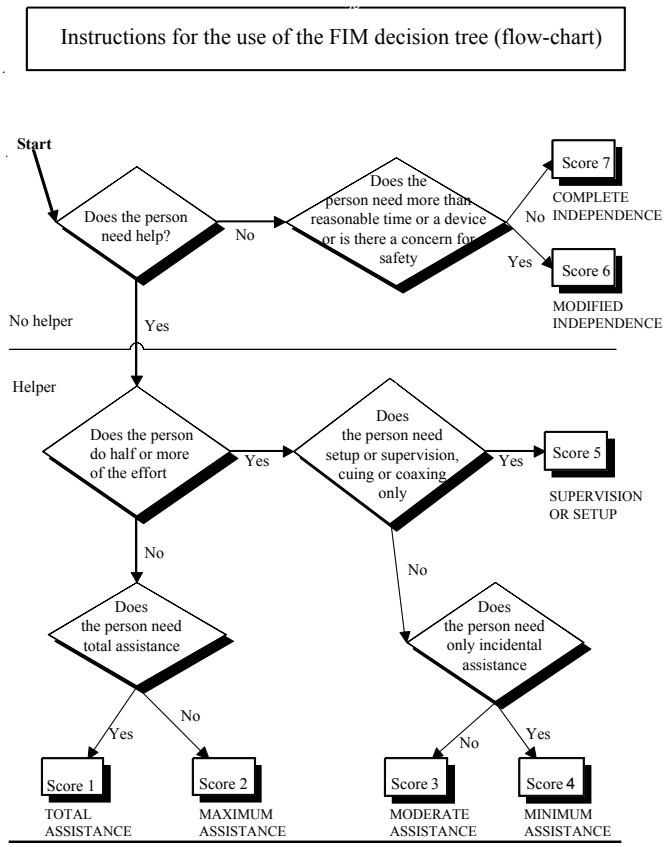

Figure 4. The FIM flow chart

A consensus score for each item was used in the analysis between the hospital assessment and the assessment made at home (Study III). In Study IV, where two different modes of data administration (postal questionnaire and interview; semi-structured according to the FIM™ and IAM instruments) were compared, the order of data collection was set, with the postal questionnaire first and (after the return of this) the interview second. The aim was to reduce the influence of any eventual thoughts provoked by the interview in the persons' questionnaire self-reports (more "naive"). All analyses made in the studies were kept blind until the end of each study (Studies I-IV). If any person needed an intervention or other information about rehabilitation needs, this was given.

Study I

The FIM™ assessments were made in home visits by one pair of raters and at the clinic by another pair of raters two years post stroke. Assessments were made independently by each of the two raters using a semi-structured procedure. Data collection started with ADL assessments according to the FIM™ independently by two raters in a semi-structured interview. The interviews were conducted in the person's home in all 68 persons by two OTs. A clinic visit could be made within a week to reproduce the interview according to the FIM™/IAM items with another pair of raters (an OT and a nurse) and included a physician's assessment. The clinical visit was completed for all but one person (n=67).

Study II

ADL assessments were made independently by a pair of raters, two years post stroke, for the eight instrumental activities of the IAM, parallel with the perceived difficulty to perform the activities. The interviews were conducted in the person's home. The raters' assessment procedure was similar to that used in Study I, but here the raters shared the same profession, i.e. both were OTs.

Study III

Paired assessments of the FIM™ and IAM instruments from two OTs´ semi-structured interviews two years post stroke were used to analyse instrumental structure and to study the dimensionality. Further, the consensus in the assessments made of FIM™ items by the two OTs was used and compared with the FIM™ observation assessments at discharge. The analyses included comparing the dependency scale with the perceived difficulty scale to find acceptable models and analyse stability over time.

Study IV

A pilot study was carried out with two persons to test the questionnaires, one with prior stroke (not a participant in the study). After the questionnaire was sent back, a revised version of the questionnaire was used in Study IV.

Study IV compared modes of administration in a special order to minimise the "carry-over effect" and the influences of the raters on self-reported ADL performance. The procedure started with the self-reported postal questionnaire, which was to be sent out first, and this data collection was required to be completed before the next stage. This ordering of the modes was assumed to minimise the influence of the rater. In the case of proxies, the person was requested to answer the question on his or her own, although it was possible to receive help with the writing if problems arose. After the return of the questionnaire, the first interview was conducted by telephone using the ADL taxonomy as a structure for the interview. When the assessment with the ADL taxonomy was completed, the postal questionnaire form of the items from the FIM™/IAM was sent out and a face-to-face interview took place after its return in a setting chosen by the participant (the home setting or clinical setting). The reason for this was that we wished to conduct the two interviews at different times, and the questionnaire had to be completed before the interview was conducted in order to minimise carry-over effects. The time between the completion of the questionnaire and the interview was conducted was one to two weeks, and the total collection process took approximately three to four weeks. In this phase after stroke (11 years) we assumed ADL behaviour to be stable and thus that the time span was acceptable.

Rater experience

The raters in the studies were experienced senior raters, either two OTs or one OT and one nurse who made independent assessments of the participant's self-report of his/her performance of daily life activities. All raters had participated in an obligatory one-day FIM training course, as recommended by the Uniform Data System, New York State University at Buffalo. The Guide for the Uniform Data Set for Medical Rehabilitation (adult FIM™ version 4.0, Swedish translation, 1994) was used in Studies I-III and the Guide for the Uniform Data Set for Medical Rehabilitation (adult FIM™ version 5.0, Swedish translation 1996) was used in Study IV.

Statistics and mathematical procedures and analysis

The ADL assessments in the studies were analysed on the item level. A description of the analysis is given in Table 6.

Table 6. Analysis used in the studies

Analysis	Study
Unweighted kappa	I, II, IV
Percentage agreement, PA	I, II, IV
Relative Operating Characteristic, ROC	I, II
Rasch analysis	III
Intraclass Correlation Coefficient, ICC	I
Wilcoxon signed rank test	I
Students t test	III
Mann-Whitney U test	III

Study I

Paired assessments of personal care, social and cognitive items in 63 participants were analysed using unweighted kappa and percentage agreement (PA value) between two raters. The assessments were made independently by each of the pairs of raters (A-B and E-Y) in a semi-structured interview. Calculations with six pair combinations of raters´assessments (A-B, E-Y, A-E, A-Y, B-E, and B) were compared with regard to the step differences and the association as well as the agreement between the two interviews within one week. The association between the raters could be established using the Intraclass Correlation Coefficient, ICC, a two-way ANOVA measure that gives the relations between assessments (76). The agreements used unweighted kappa statistics. Step differences were analysed to identify the most common categories that caused disagreements in the estimates of the participants' problem areas. Bland-Altman plots were used to study rater differences according to the physical and social-cognitive items of the FIM™. The cumulative frequencies illustrated the systematic differences in the ROC curves between raters in their use of the different categories of the scale.

Study II

Paired assessments of instrumental activity items in 63 participants were analysed using kappa agreement, percentage agreement, PA, value and cumulative frequencies (%). The assessments were made independently by two different raters in a semi-structured interview. The cumulative frequencies illustrate the systematic differences between raters in their use of the different categories of the scale.

Study III

Rasch analysis was used to analyse structure and hierarchical orders of the items in determining "item difficulty" and "person measures" concerning the construct validity of the FIM™ and IAM.
The "item difficulty" and the "person measure" are result of the calibration with Rasch analysis to transform the assessments to achieve interval data.
The study focused on the structure of the instruments and the stability of the items, e.g. the changes in ranking order in different environmental settings. Additional Rasch analyses were also performed to compare the "person measures" made by the pairs of raters.

Study IV

Kappa statistics and percentage agreement, PA, were used to compare two modes of self-reported ADL ability in P-ADL and I-ADL items, in a postal questionnaire and a semi-structured interview, with the items from the FIM™/IAM and the ADL taxonomy. The study focused on modes and analysed the results of self-reported paired ADL items according to in/dependent performance. All data collected on individual ADL in/dependence were dichotomised to an "independency category" and a "dependency on another person category" in both instruments. For the seven categories of the FIM and IAM, categories 1 to 5 were collapsed into one dependency category and 6 to 7 were collapsed into an independency category. In the ADL taxonomy, this collapse gave a dependency score using categories B to F. Category A was kept for the assessments of independency. Category Ö was assessed as dependent. The dichotomised results of independent and dependent ADL performance were compared in the items from the FIM™/IAM instruments and the ADL taxonomy.

The unweighted kappa coefficient

The coefficient kappa is the most accepted measure of agreement concerning data from nominal and ordinal scales (84) and was introduced by Cohen in 1960 (15). The kappa measure is a chance-corrected, scaled agreement measure and is appropriate for use with nominal or ordinal data (Figure 1). Kappa is defined by a relation between the observed proportion of agreement and the expected proportion of agreement by chance (1). Unweighted kappa is an analysis of exact agreement, "that is it treats agreement as an all or none phenomenon with no room for 'close agreement", i.e. the raters

use the same category for each participant assessed (71). The best and most informative analysis is achieved if kappa values are computed for pairs of raters on the item level. Unweighted kappa will not indicate whether most of the disagreements are accounted for by one specific category or rater. Kappa statistics was chosen for its correctness concerning rater agreement. The studies focus on each item and the category levels of the instruments. However, a kappa value does not differentiate between disagreements (71).

Values for a kappa coefficient exceeding 0.40 (0.40-0.75) are considered fair to good, and those exceeding 0.75 are considered to have excellent agreement, while values below 0.40 are poor according to Fleiss (29, p. 218). In Altman (1) the cut-off point is somewhat different, showing moderate to good agreement between 0.40 and 0.80 and very good agreement above values of 0.81 (55). A p-value of <0.05 was considered statistically significant.

Table 7. Strength of the kappa agreement *

Value of *k*	Strength of agreement
<0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very good

* (55)

Percentage agreement, PA

The marginal distributions and disagreements between the raters were studied in the contingency tables (Figure 5). Percentage agreement, PA, was used to determine the number of exact agreements between the raters. Exact agreements can be seen in the diagonals in the contingency tables (Figure 1). Good percentage agreement can be expected to exceed 80% (52).

Cumulative relative frequency

The cumulative relative frequency for each rater was calculated from the marginal distribution of each rater, showing the rater´s use of the seven-step scale in the FIM™ and IAM (Figure 5). When there are skewed marginal distributions and the raters use only part of the ordinal scale, as in Studies I and II, kappa values can be low compared to the PA values (27). The PA value can be the same but gives a range of different kappa values depending on the marginal distribution.

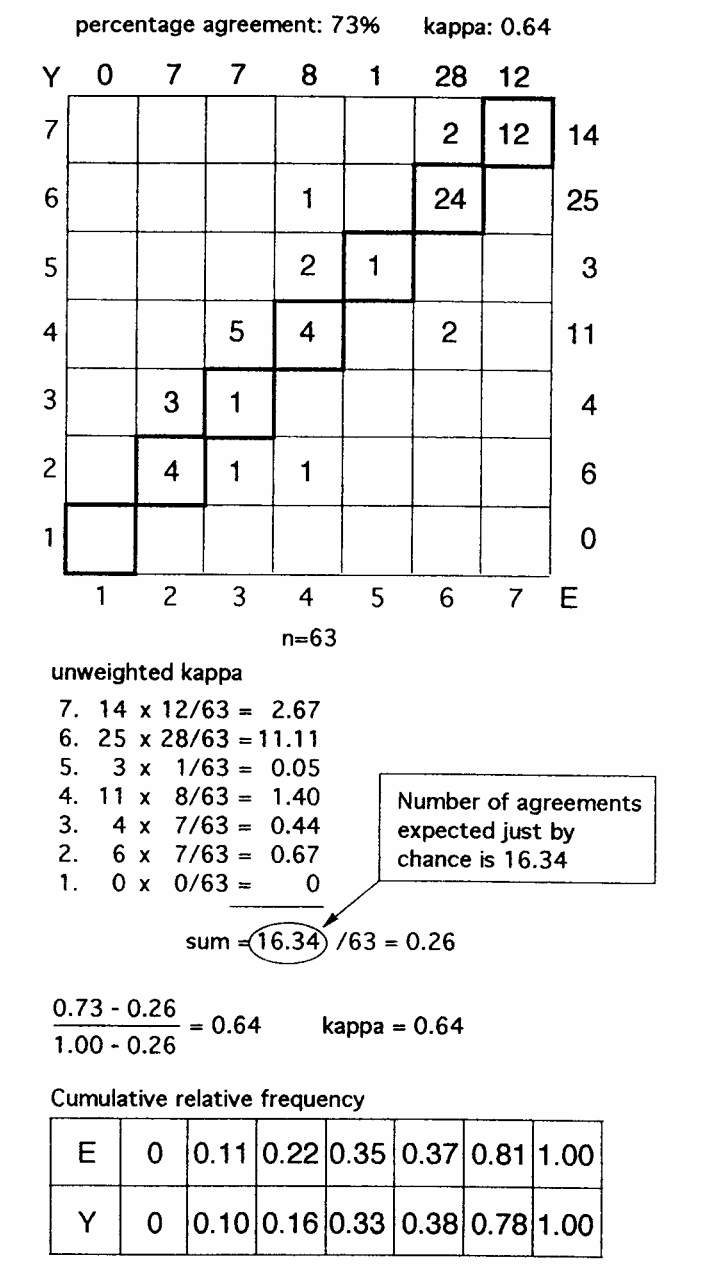

Figure 5. Example of a contingency table, percentage agreement, PA, kappa and the cumulative relative frequencies

Systematic differences - ROC curves (not included in the publications)

In contingency tables any different use of the scale categories between the raters will be shown as divergences from the diagonal (Figure 5). The dispersed observations from the diagonals in the contingency table are a sign of random or occasional disagreements. Further, the systematic differences (bias) between the raters can be calculated from the cumulative frequencies for each item and each category (Figure 5). The cumulative relative frequencies were plotted in a Relative Operating Characteristic, ROC curve (Figure 6) for each FIM™ and IAM item (86), (Figure 2). If one rater consistently underestimates or overestimates relative to the other, the ROC curve will be located to one side of the diagonal of agreement, e.g. systematic disagreements; this can be seen as a concave or a convex curve. This might occur when the raters use a rating scale that has a basis in

personality and/or professional differences (8, 84) (Figure 6) or different interpretations of the scale categories. In addition, both occasional and systematic disagreements in the items can be calculated, each with specific values according to a non-parametric statistical method (85).

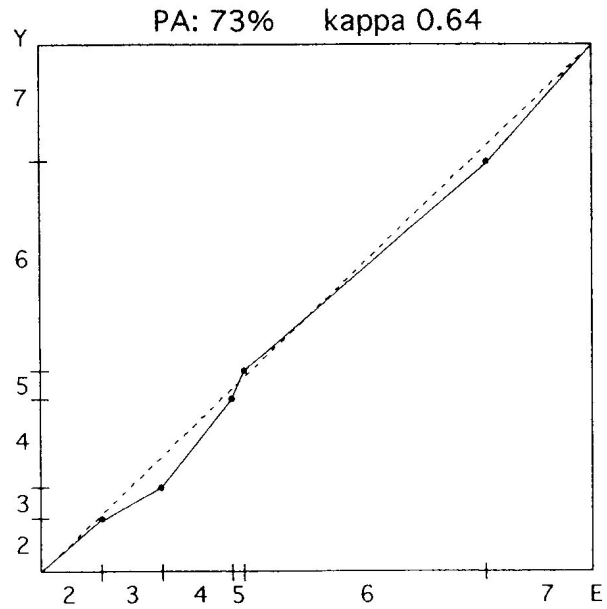

Figure 6. The cumulative relative frequencies plotted in a ROC curve. It shows that one rater (E) more frequently uses the lower categories.

Rasch analysis

The conventional unit for Rasch analysis is logit value (log-odds units), and the centre of the scale is set at 0. The analysis provides fit statistics, i.e. how well different items describe the group of persons, and can also be used to examine how well an individual fits the whole group. When an item does not perform as expected, the fit statistics flag an unexpected behaviour of an item or an unusual person.

The Rasch model transforms the ordinal data into linear measures by using the relative ability of the persons and the relative difficulty of the items. The item measures will be placed in ranking order from easy to difficult. The "person measures" will be arranged from "most able" (independent) to "less able" (dependent) on a linear scale. The physical and social-cognitive items were separated in the analysis. Further, the social and cognitive items were separated into two sub-groups of participants depending on whether there was any linguistic problem, "aphasia" or "no aphasia". The Rasch analysis can be viewed as a construct validation of instruments by assessing the homogeneity of items under different conditions and identifying person measures and misfitting items.

Rasch analysis was done using a software program for PC (BIGSTEPS) (98).

Intraclass Correlation Coefficient, ICC

The Intraclass correlation coefficient, ICC, has been widely used to analyse assessments between more than two raters. As ICC is a measure of the degree of "association between two quantities, it does not measure how closely they agree" (1) (4). However, an inappropriate use of correlation coefficients may give a misleading analysis and incorrect information about agreement and reproducibility (75). There are different ways of calculating the ICC values depending on the procedure used to collect data (76), either using a one-way ANOVA or a two-way ANOVA. In this study, the two-way ANOVA is used since the design (3.1) of the study was that "each target is rated by each of the same k judges, who are the only judges of interest" (76). The ICC is equivalent to weighted kappa statistics since the assessments are weighted and small or close divergences are accepted (71).

The recommended coefficient varies depending on whether the purpose is for group comparisons (0.60 – 0.70) or for making decisions about individuals (0.90 or more) (70). Ottenbacher also emphasises different recommendations for the ICC values in his review of FIM™ reports (68). In comparing groups, coefficients of 0.75 and above are "indicative of good reliability and those < 0.75, poor or moderate reliability" and the author recommended values of at least 0.85 in making decisions about individuals.

Wilcoxon signed rank test

Significant differences between the raters on the item level were analysed with the Wilcoxon signed rank non parametric test. A p-value of 0.05 was considered statistically significant. The Bonferroni-Holm method (44) was applied for correction of the results from mass significance.

T test

An ordinary Student t test was used in Study III to calculate differences between the two occasions. After transforming the ordinal data into linear data with Rasch analysis, the parametric t test can be used for calculation of individual differences in the measurement values (logits).

Mann-Whitney U test

The differences between groups were tested with the Mann-Whitney U test according to gender and age (Study III).

Ethical considerations

All participants gave their written and verbal informed consent for each of the studies after both written information was given and verbal information was given by telephone. All data were

collected before the analysis started. The Ethics Committee of the Faculty of Medicine, University of Gothenburg, Sweden, approved the different studies.

RESULTS

Study I

The reliability of an interview approach using the FIM™ instrument showed, on the item level, a generally high inter-rater agreement during the same interview independent of the pair of raters. However, the consistency of the ADL assessment over time made within the interval of one week showed decreased stability according to the kappa values as well as when analysed with the Wilcoxon signed rank non-parametric test. The Wilcoxon signed rank test showed significant differences between the raters in the items: *Dressing upper body* and *Dressing lower body*, *Transfer toilet*, *Transfer tub/shower*, *Walk/wheelchair* and *Social-cognitive*. Significant differences could be seen in analysing differences between each pair of raters in the two contexts; significant differences could be seen, except for the pair of raters at the clinic, in physical items. In the home context, there was a small significant difference between the raters also in assessing the physical items of the FIM™ (Study 1). In the Bland-Altman plot (Study I) comparisons could be made of the domain of items (physical and social-cognitive) on the individual level. This analysis generally showed a similar pattern of disagreements independent of the individual severity of the disability (sum score) (high or low degree of dependency on another person).

The most frequent category differences, independent of items, were, for both pairs of raters, A-B (E-Y), in determining between the two independence categories: “complete independence” and “modified independence”; 32 % (28 %). There were some further differences between the pair of raters indicating greater difficulty for the pair to be in agreement in the hospital setting in the categories of “modified independence” and “supervision” 22 % (14%) and at home in the categories of “supervision” and “minimal contact assistance/occasional assistance” 16 % (17%) (Study I). In general, the transfer and social-cognitive items caused more disagreement on the item level for both pairs of raters.

Analyses of the cumulative relative frequencies for each FIM™ item were calculated in ROC curves, and the systematic differences between the pair of raters (at the clinic and in the home environment) were studied (Figure 7). In general the systematic differences for the 13 physical items of FIM™ were relatively small, as can be seen in the ROC curves, but there were some discrepancies. The discrepancies were seen more often in the transfer items and in the social and cognitive items. It can be noticed that the assessments are generally located in the upper part of the scale, e.g. more independent persons. Figure 7 gives two examples of significant differences in the items of *Eating* and *Transfer toilet* that caused more problems for the raters making the assessment in the home setting. The raters had greater difficulty in *Eating* to determine whether the participant

was independent or not (category 6 or 5, independent or not) and concerning *Transfer toilet* category 7 and 6 (modified independence or complete independence). *Social interaction* showed systemic differences in both assessment settings.

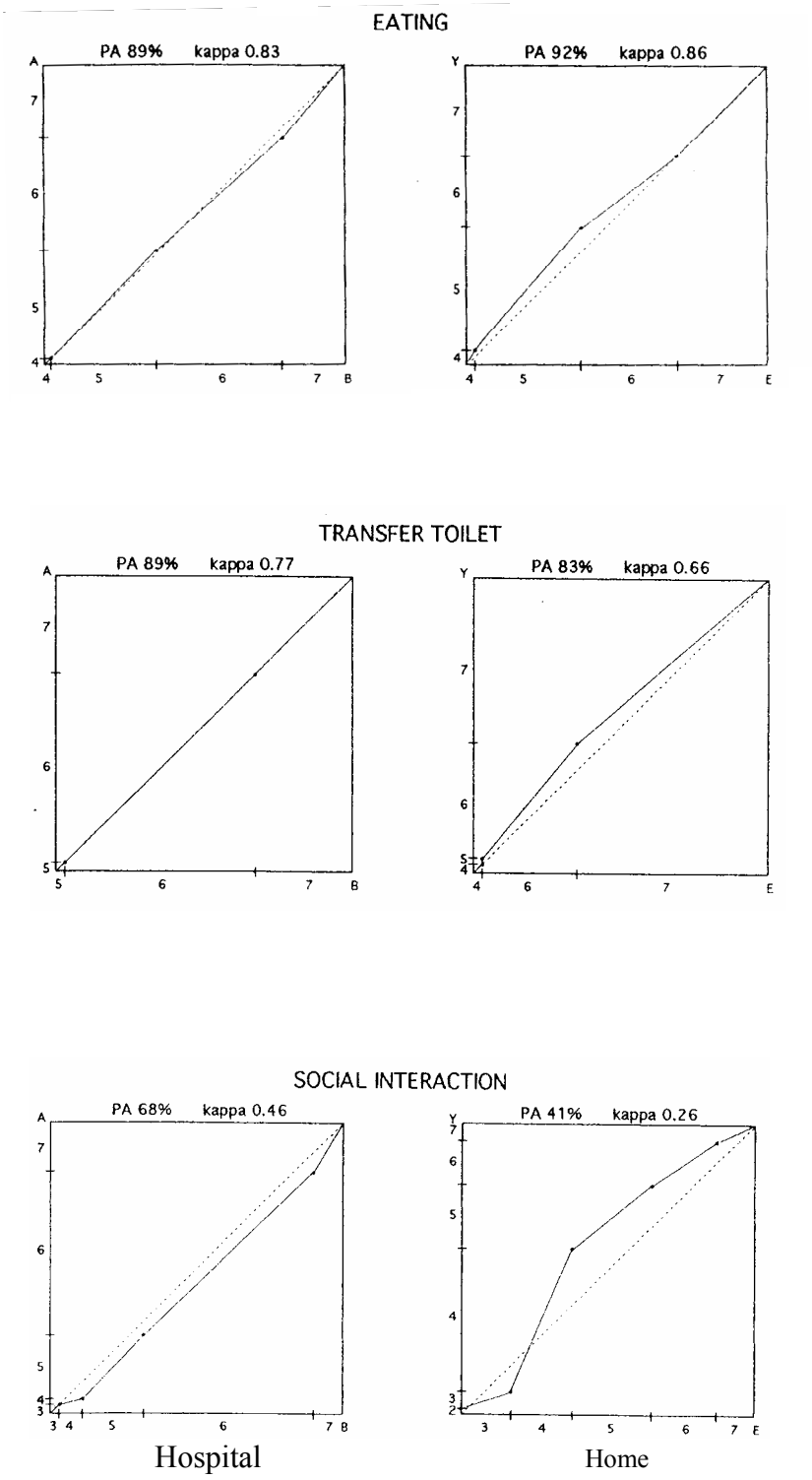

Figure 7. The ROC curves, indicating systematic differences between the pair of raters for (E-Y, home context) in all three examples, but only in *Social interaction* for (A-B, hospital context.

Study II

According to the kappa values, the inter-rater agreement in the IAM instrument had good reliability. All eight IAM items had kappa values exceeding 0.60 with a range between 0.63 and 0.75. Corresponding PA values had a range between 70 and 83% (Table 8).

However, the cumulative frequencies showed inconsistencies between the raters in using the seven categories that assess common instrumental activities. Level 5 was not used by rater Y for *Small-scale shopping* or by rater E for *Cleaning*, while level 3 was not used by rater Y for *Transportation* and level 1 was not used at all in *Locomotion outdoors* (Study II, Figure 2).

Table 8. Kappa and percentage agreement (PA) presented for raters E and Y according to the eight IAM items.

IAM items	Raters E-Y PA (%)	Raters E-Y Kappa
locomotion outdoors	73	0.64
simple meal	78	0.71
cooking	79	0.73
public transportation	83	0.75
small-scale shopping	75	0.68
large-scale shopping	76	0.69
cleaning	70	0.63
washing	78	0.69

The steps in the upper part of the IAM scale (assessing less dependency on another person) were used predominantly for ratings of *Locomotion outdoors, Simple meal* and *Small-scale shopping*. For the remaining items, *Cooking, Transportation*, *Large-scale shopping*, *Cleaning* and *Washing,* the more dependent categories were used in approximately 50% of the assessments by both raters (Study II, Figure 2).

There were also similarities in the IAM and the FIM™ ratings concerning the difficulty in determining whether a "complete" or "modified independency" category is appropriate. The most frequent differences (16 %) were in the independent categories in distinguishing whether the respondent was "modified independent" or "completely independent". There appeared to be difficulties in all the instrumental items, as 23 % of the assessments caused problems in the decision about whether the patient was dependent or not: (7-4), (6-5), (6-4), (6-3), (6-1). Two category differences or more (≥2) caused a further 14% disagreement. In approximately a quarter of the total assessments (63x8), there was some kind of inconsistency between the raters.

In analysing the systematic differences between the raters, the cumulative relative frequencies were plotted in an ROC curve for each IAM item (Figure 8). As can be seen, there were no great deviations from the diagonals of the ROC curves. This emphasises the random or occasional

influence on the scorings being the most important reason for the inconsistencies between the two raters.

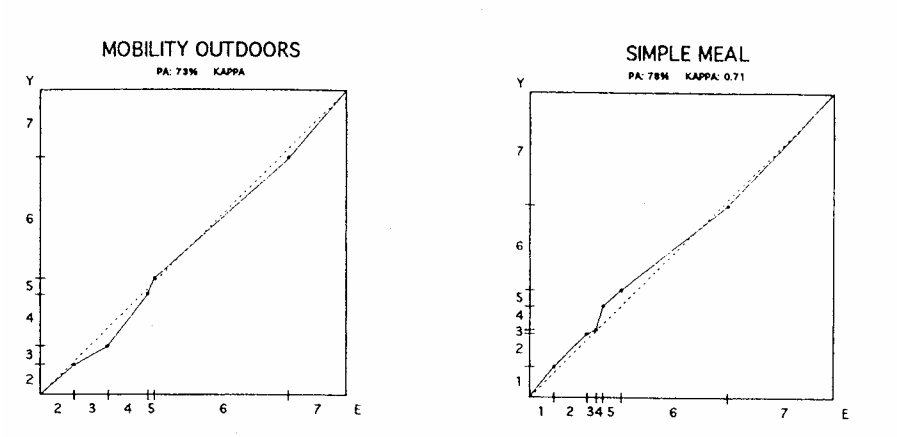

Figure 8. ROC curves showing an example of the assessments for one pair of raters of two items in IAM in the context of the person's home environment.

Study III

It was possible to combine the ratings of dependence in physical items in FIM™ and instrumental items in the IAM in a calibration for all items according to Rasch analysis (98). However, the instruments' dimensionality and structure were analysed according to difficulty of items and in/dependent performance. The ranked order of the items of the FIM™ and the IAM could identify different domains. The items could be arranged in different areas, e.g. self-care, mobility indoors and instrumental items. According to Rasch analysis, the instrumental items were in general clustered in two areas with different difficulty of the items: the easier items (*Simple meal*, *Locomotion outdoors* and *Small-scale shopping*) and the items in which it is more difficult to achieve high independence (e.g. more dependent persons) such as *Large-scale shopping*, *Washing*, *Cleaning*, *Public transportation* and *Cooking*. The hierarchical orders of the motor items showed that the transfer items of *Transfer toilet* and *Transfer bed/chair*, as well as *Grooming* and *Eating,* were the easiest of the physical items (Study III, Figure 5).

It was also possible to obtain an acceptable model for a calibration of the FIM™ activities of physical items and social and cognitive items, but separate domains, for discharge and follow-up, with a reliability of 0.94. However, the study showed changes in the order of severity for several of the personal care items of the FIM™ and all the social and cognitive items (Study III, Figure 2) compared to the clinical setting. The results showed a significantly increased use of the dependency categories for personal care items, among others in *Eating*, *Dressing upper* and *lower body*, *Shower/Bath* and *Bladder* (Study III, Figure 1).

The "person measure" logit values (Study III, Figure 3) derived from the Rasch analysis of the physical items of the FIM™ also showed a significant increase in dependence on another person at follow-up, in 51 % of the individuals, and a significant decrease, in 10 % of the individuals (n=68). The "person measure" values for the social and cognitive items also showed a significant increase in dependence in 53 % of the non-aphasic persons and 64 % of the aphasic. The reduction was 5 % and 4 %, respectively (Study III, Figure 4). The item calibration (how well the different items describe the group of persons, shown on a linear scale) during inpatient care was studied in an earlier report by the same department (37) that confirmed a stable use of the order of FIM™ ratings at admission and discharge in the hospital environment (the same hierarchical order) and showed significant changes and improvements in most of the patients between admission and discharge.

A combination of physical activities from the FIM™ and IAM also gave acceptable models for both dependence and perceived difficulty. In general, there was agreement between the ratings of dependence and perceived difficulty, but some discrepancies were noted, e.g. in *Locomotion outdoors* and *Public transportation* (Study III, Figure 7). There were also differences in dependency with respect to gender and age. Men found it more difficult to be independent in such instrumental activities as *Cooking, Cleaning* and *Washing* than women; the opposite was true for *Small-scale* and *Large-scale shopping* and *Locomotion outdoors* (Study III, Figure 8). Persons aged >55 years had a slightly higher level of dependence and perceived difficulty in IAM activities than those below that age.

Study IV

The comparison of self-reported data between the modes of postal questionnaire and interview generally showed moderate to good agreement and significant kappa values in the ADL instruments, with some exceptions. These items had fair kappa values, e.g. ≤ 0.40. The items from the FIM™ were: *Bladder*, *Stairs*, *Social interaction* and *Memory* (Study IV, Table 2). The items from the IAM were *Cooking*, *Large-scale shopping* and *Public transportation*, and the item from the ADL taxonomy was *Going to the toilet*.

Non significant kappa values were found in the FIM™ items of *Grooming*, *Toileting, Transfer toilet* and *Problem solving.* Non significant kappa values in the ADL taxonomy were in the item of *Eating and drinking*. All items in the IAM showed significant low kappa values (range 0.36-0.47) except in the case of *Small-scale shopping* (0.80). The simpler activities, such as *Locomotion outdoors, Simple meal* and *Small-scale shopping,* showed a high degree of independence in each of the items as compared to the other five items, regardless of mode, with significant, but moderate, kappa statistics (Study IV, Table 3). A disrupted rank ordering of actions (labelled "Ö") was found in both the questionnaires and the telephone interview using the ADL taxonomy. The disrupted rank order in all assessments (n=36) identified in the frequency analysis was most common in the more

complex and socially influenced items, such as *Transportation* (n=20), *Shopping* (n=7) and *Washing* (n=10). These actions were assessed as dependent.

The results indicate problems in some items, different in each ADL instrument, in reproducing the self-report and show that further analysis is needed. The FIM™ instrument demonstrated the greatest problems in *Grooming* and *Transfer toilet,* and the IAM with *Cooking* and *Public transportation.* The most problematic items in the ADL taxonomy were *Eating/drinking* and *Going to the toilet* In general, the kappa values of the ADL taxonomy ranged between 0.50 and 0.80, except *Eating and drinking* (0.14) and *Going to the toilet* (0.35), which showed a somewhat higher level of stability (agreement) with the reproduced self-report compared to the FIM™/IAM.

Further comparisons between the two ADL interviews (the FIM™/IAM and the ADL taxonomy) generally showed good agreement according to how well the assessment tools and mode classified dependence in ADL performance. In the interview, most of the items of the personal care identified ADL dependence regardless of the tool. ADL dependence in personal care items was found in 16 persons (44%) in the items in the FIM™ and in 21 persons (58) % in the ADL taxonomy. No one in the study group was assessed as being fully independent in either of the two I-ADL indices (IAM/ADL taxonomy). The most independent persons (n=8/n=10), five of whom were the same in both instruments, were all assisted in one or two instrumental activities such as *Cooking, Large-scale shopping*, *Cleaning* and *Washing* or did not carry out some of the actions in *Transportation* (ADL taxonomy).

Limitation of the studies

The selected sample is a convenience population from the rehabilitation ward, consecutively included in the two-year follow-up over a period of two years. The patients´ functional abilities during a rehabilitation period after stroke are heterogeneous and include a broad range of patients, from recovered ability to activity limitations and, at the extreme, a need of support in all daily activities. The participants selected for the studies lived in the community, and the aim of the study was to examine the stability in both personal care and instrumental activities. For these reasons, floor and ceiling effects emerged in the personal care items. This skewness in the sample makes it difficult to obtain significant kappa values in certain items, and this is a weakness in this community living sample.

The aim of Study III was to assess the difficulties on the item level using Rasch analysis, with data gathered at discharge and at a two-year follow-up. Despite the use of “decision trees” to assess dependent performance, it is not possible to control for the influence of two different modes of data collection (observations and interviews); neither is this possible for the two environments or the time factor (changes in stroke impairment). Therefore, in Study III, it was possible to identify different difficulties on an item level, although the statistics and models used do not explain the

cause of the variation in step categories. It is also a limitation that it is not possible to distinguish the intra-rater assessment error from the intra-subject inconsistent performance.

The kappa value is limited when there is a low variation in the results, as in these studies in the items of personal care, as a result of a ceiling effect. Here, most of the patients were independent in most of the P-ADL activities (4, 22). There is also a limitation in Studies I and II in that the kappa values were not analysed according to significance levels. Study IV showed that the kappa values gave limited information, as there was a skewed distribution among the categories, with a limited use of some categories.

DISCUSSION

The overall aim of the studies was to analyse the stability of the assessment of self-reported ADL after stroke. The assessed construct was ADL in/dependence and a performance approach was used. The outcomes of the assessments were analysed between different raters, instruments and modes of administration. The ADL assessments were made with regard to both personal and instrumental items. A semi-structured assessment procedure was used with raters in different situations and environments and the interview method was compared to a self-reported questionnaire.

The studies generally showed a moderate to good level of agreement showing stability in the ADL assessments in several of the items, regardless of rater, instruments or modes of administration in assessing personal and instrumental items. A high level of agreement is a condition for following real individual changes over time and thus minimising random and systematic disagreements (84). The instability and variability in the assessments is a reflection of random or systematic disagreements.

Impact of stroke on ADL assessments

The impact of stroke shows an individual time course and different severity depending on the person. The health condition after a stroke might result in different ability over a 24-hour period (73, 77). Fluctuations in the ability to perform daily activities that are caused by the stroke might always exist and thus complicate the assessment (61, 73, 77, 88). Fluctuations in the ability to perform daily activities that are caused by the stroke might always exist and thus complicate the assessment (61, 88).

The ability to perform daily activities late after a stroke, as in the present studies, is not expected to be influenced to a great extent by the health condition after the stroke. Co- morbidity was seen in approximately one-third of the group, in whom cardiovascular problems were the most common, followed by musculoskeletal problems (n=5). Further medical conditions were not assumed to influence ADL to a great extent. The ability to perform depends to some extent on the specific

situation in which the activities are performed, where different capacities are needed. ADL and the problems that are found are integrated and are influenced by different environmental conditions, such as social- attitudinal and physical aspects (96).

The different solutions to the individual ADL problems were influenced among other things by the cognitive understanding of the impact of the stroke (awareness of the consequences) and the individual's personality (for example asking for assistance or carrying out the activity with for example reduced quality). The deficits in emotional and mental function may even be apparent in persons who have experienced a relatively mild stroke, where there may be symptoms such as depression, tiredness and limited concentration and physical and mental endurance, which may increase the variability in the performance in daily life and the dependence on another person (9). It is also difficult to distinguish a normal general fluctuation in activity performance because of different habits and roles, as is seen in the normal population (79) in the volatility of the performance caused by the stroke. At the same time, professionals must be able to handle all kinds of stroke patients and their ADL situations.

The influence of raters

Fluctuations in the ability to perform activities are a problem for raters. Raters must interpret and transform a general behaviour according to guidelines to a specific scoring category. The intra-rater "errors" are difficult to separate from subjects´ in assessing ADL performance (22).

The inter-rater agreement in the same interview setting was generally high (Studies I and II). The reliability of the FIM™ physical items has been confirmed in several studies (39, 68, 83)in contrast to the more problematic social and cognitive items. The latter items generally showed less inter-rater agreement in all items (23, 78). The personal care items showed greater agreement than the instrumental activity items (72), as was found in the present study (Study I).

The semi-structured FIM™ interview held within a week of a reproduced assessment by another pair of raters showed only moderate to low agreement according to the items' kappa values, above all in *Transfers*, *Locomotion* and *Social-cognition* items (Study I). Only one item had an ICC value above 0.85: *Shower/Bath*. The inconsistency between the assessments was more obvious in the transfer items. This may be related to the ceiling effect, as most of the individuals were independent and kappa values are sensitive when only a part of the scoring categories are used (27). Contradictory results were found concerning the physical items in a cross-validated study between countries in hospital settings, where the transfer items showed fewer disordered categories (thresholds) (61). The lesser concordance between the two interview occasions emphasises that there are some problems in the assessments, such as the concept of "in/dependent performance" (58, 61, 88).

The most problematic category in both interview settings (clinical and home) arose when the raters were to assess and choose between the two categories having to do with independence: "complete independence" or "modified independence". There was some further disagreement between the interviews: the raters that conducted the interview in the home had somewhat more difficulty in deciding between the categories of "supervision/set-up" and "minimal or incidental assistance". The raters in the clinic setting had somewhat greater problems in deciding whether the person was "modified independent" or needed "supervision" (Study I).

The instrumental items of the IAM showed good agreement (Study II) according to the kappa and PA values between the two raters during the same interview. The combination of kappa values and percentage agreement was useful in understanding the disagreements between raters when all categories were used. Thus, a high PA value and low kappa value could arise when the raters only used a limited part of the scale (27). The disagreement between the raters emerged in general in a quarter of the assessments in all eight items, independent of the type of item (Study II, Table 1). This has been confirmed by other researchers who have shown the complexity of assessing instrumental activities (20, 58, 72, 82).

Systematic differences were evident when one rater consistently interpreted the step categories as higher than or lower than the other rater. In the FIM™ analysis, more systematic differences were seen in the social and cognitive items by both pairs of raters, as compared to what was observed in the physical items. There were some systematic differences in all of the items of the FIM™ and IAM. Further analyses, for example of a distinction between random and systematic disagreements, were not made in the present studies. It was possible that the random differences hid true systematic changes (84). In the next step, in a review the reliability of methods, a method can be to used to distinguish between the random and systematic the disagreements (87). The rater-related variability in the interview method may be controlled to some extent by training in the instrument. However, there are aspects that were not investigated in the present study, such as the raters' cultural context. In the present studies, it seemed that the experiences and the context of the raters in their professions might also have an influence on the assessments and that there may also be an effect of training (Study I) (30). In the author's understanding, the most reliable ratings with the FIM™ were made by clinically active occupational therapists, with no prior experience of the instrument before their training (30). However, in another study that compared ADL assessments made by physicians and non-medical professionals, where the latter received training in the Barthel index (74), the trained non medical professionals scored as highly as the physician, indicating that training does matter.

The influence of the instrumental structures

The ADL instruments selected (FIM™/IAM and the ADL taxonomy) were generally suitable for describing the ability to perform ADL, and the most independent and most dependent persons were

identified. Most of the persons who were dependent in the more simple instrumental activities, such as *Locomotion outdoors* and *Simple meal,* were identified in all of the instruments. The majority were the same persons. However, there were differences between the instruments with regard to the concept of the definitions of the items.

The description of the individual ability to perform daily activities depends on the definition of the items in the ADL instrument and the item severity. The lack of a consensus on a "golden ADL instrument" is a problem in the comparison across countries of stroke outcome research studies.

The hierarchical order of the items in the FIM™ and IAM changed between the assessments made at hospital discharge and those made in the home setting, despite relatively stable stroke (Study III). The findings showed a relatively greater dependence on assistance from another person at the two-year follow-up. These results were also found in another stroke study (31). The difficulty of the items changed, to be more difficult in the home environment, especially in the items of *Eating, Bladder*, *Grooming* and *Transfer shower/bath* (Study III, Figure 1). A comparative study in the FIM™ used in different countries showed invalid items in the comparisons, especially in the items of *Eating* and *Bladder* (62). The ranking order of the items in the FIM™ has also been reported to be different in comparisons in hospital settings in different countries (61). This makes it difficult to follow individual ADL ability over time, as the validity of the assessments depends on the environment. The ways in which environmental circumstances influence ADL assessments have not been considered at great length in the literature, even though they clearly influence assessments of functional abilities (58). The way in which environmental aspects can contribute to decreasing the reliability of the instruments used needs further research.

The common disagreements between raters in the categories of "modified independence" and "supervision" or "minimal assistance/incidental assistance" are obvious (Figure 4). This might be related to a situational or temporary dependency after a stroke, where ADL behaviour is insufficiently transformed and described in the ADL instruments, as was confirmed in another study (14). The fluctuation in the ability is to some extent due to the impact of the stroke, which increases the demands on the ADL instrument and method that will be used in the stroke population. The instruments need to more clearly capture this phenomenon of a situational or temporary assistance of another person depending on different situations, as was also pointed out in a study that compared the FIM™ and Sunnås ADL index (14). There are several reasons why persons with stroke need assistance in certain situations, depending on the physical environment: there are barriers in getting on the toilet, time limits, needs for special clothes, fatigue or other situational demands. The way in which this situational dependency can be clarified in the description of the ordinal categories must be further studied. That this particular ADL situation arose in one interview but not in another might be because the semi-structured interviews permitted some different questions and interpretations, exemplified as "step differences" between pairs of raters seen in Study I, Table 2.

The FIM™/IAM and ADL taxonomies are built on different concepts. The activities/items have different operational definitions and response categories. In general, the FIM™ instrument described all transfer items separate from other activities. For example, an individual with greater disability might be independent in *Toileting transfer* but need assistance with clothes or hygienic tasks according to the FIM™. In the home environment, it might be impractical or even impossible to distinguish between these two needs of assistance. The response categories of the FIM™/ IAM have seven ordinal categories, compared to the two categories in the ADL taxonomy: "does perform" or "does not perform" each action. However, the actions in each activity in the ADL taxonomy are ordinal in their structure (82). Differences in these rank orderings, indicating problems with the concept and the reasons for not performing the activities, might have different causes other than the stroke impact, and may for example have to do with the influence of interests and roles of the person. A problem in the ADL taxonomy might be that it contains a great many items (n=47), However, it is possible to reduce the items to suit the aim of the assessment in correspondence to the guidelines in order to suit persons affected by fatigue and other cognitive deficits after a stroke.

Comparisons between the ADL instruments, items from the FIM™/IAM and the ADL taxonomies showed some differences between the items reported. This can be explained by the different operational descriptions of the items in the instruments. The ADL taxonomy includes for example the actions of *Manicure* and *Pedicure,* which include difficult grip abilities that are not assessed in the FIM™.

The relatively little use of the middle categories (scale categories 3 to 5) reveal problems in using several categories to assess ADL dependence; this has also been emphasised in other studies (14, 34, 66). Recent studies have also analysed the FIM™ items using the seven ordinal category scale to assess independence, and different suggestions have been offered for reducing the scoring levels to five category levels (14, 18, 34, 37, 66) or to three scoring categories without "loss of information" (18). Independent of the method used, a reduction of the seven ordinal categories in the FIM and IAM instruments might increase the level of stability of the assessments and improve cross-cultural comparisons (18, 66) . This kind of change may possibly increase the agreement of the ADL assessment when used by different raters. The number of items in the instruments could possibly be reduced in medical settings, as several of the more complex and difficult instrumental activities are not focused upon in interventions.

The influence of the assessed concept: In/dependence

The studies showed difficulties in recapitulating ADL independence using only "doing performance" that were not further specified. This indistinctness in the concept of independence has an impact on the reliability of the ADL assessments. The ratings of dependence and the need of

assistive devices have to do with the situation, the context or the difficulty perceived by the participant. ADL performance in ordinary life comprises a long continuum from "can do", "being able to do if obligated to", "doing the activity with another person", "actually doing" or just "doing the activity", not further specified. This shows the unbounded, invisible transition between independent levels and dependent levels in real life that are seemingly more complex in community settings than in hospital settings (13, 88). There might be a need to limit the focus of the continuum of in performance approach to activities of daily living. A more distinct focus in assessing ADL independence in terms of "is able to do" might make the situation clearer. The underlying latent trait of the ability to perform activities, e.g. the continuum between "is able to do", "is able to do if needed" and "co-operates with another person", needs further study in how to control and express this in the assessment procedure. Further analyses of the ADL assessment procedure involving both personal and instrumental items could, with advantage, be guided by the ICF model and its structure (96).

Despite dependence on another person in several instrumental items, many of the participants were able to perform the activity if they were obligated to, for example if they were alone for periods, indicating another influence than stroke, such as interest and habits. These results are confirmed by another study, where instrumental activities seemed to be more changeable and influenced by physical and contextual conditions (46).

It is also important to identify persons at risk of developing inactivity and a reduction of activities and participation. In addition to the generic ADL assessments, a more client-centred assessment, such as COPM, might bring deeper insight into planning and evaluating individual interventions.

The way in which the concept of independence is categorised and analysed is also important. An example might be a study in which another type of ADL scale was used in which the responses were dichotomized into "receives no assistance" and "receives assistance" (28). This study concluded that activities of daily living (except some instrumental activities) were stable across administration periods and across settings in a large cohort of elderly persons, >85 years, in contrast to the results of the present study.

The influence of the mode of administration

The present study showed some significant differences between both pairs of raters, pointing out different questions and other situation-related information that emerged in the semi-structured interviews (Study I, Figure 2 a and b). The contexts in which professional work (in hospital or in the home context) might also influence the ADL questions as the assessments in hospital are in general focused more on capacity. This might explain the differences in the interpretations of ADL ability and their transformation into different categories in the instrument.

A generally moderate to good level of stability (agreement) was noted between the modes of administration in most items in the three ADL instruments, although some items varied (Study IV, Table 2, and Study IV, Table 4). The use of the ADL taxonomy generally showed greater stability in the items, regardless of the different modes, compared to the use of the FIM and IAM instruments. Several of the instrumental items in the IAM showed problems, with a poorer agreement between different modes of data collection (Study IV).

Self-perceived ability is a basic condition for the intervention programme and, as Law pointed out, self-perceived health must be identified as important, as it influences occupational performance after stroke (58). The approach of self-perceived ability has been shown to be more reliable than proxies, as proxies tend to overestimate problems (25, 26, 43). Other studies have also found self-reporting to be reliable (48, 49, 78). This shows the possibility of using self-reported questionnaires in the clinical setting, which might increase their clinical utility, i.e. in saving the time and energy of both the patient and the therapist. One study showed that the inter-rater reliability of the Barthel index, regardless of the mode used (postal, telephone or personal interview in the home), was high according to weighted kappa statistics in the personal care items (42). However, the modes are not interchangeable for the same individuals. This agreed with a study that compared telephone and personal interviews concerning personal care items of the FIM™ (78). A study that compared a postal questionnaire with an interview found, in agreement with the present study, that postal questionnaires were a satisfactory alternative to direct administration, except in the case of some instrumental items (10). However, the two methods were not found to be interchangeable in monitoring the same patient (10, 24). Self-reported ADL performance in the semi-structured interviews varied between different raters and situations. The superiority of one method in gathering information on ADL over the other is still unclear and probably varies depending on the aim. In general, agreement could be maintained well by a change of method to a postal questionnaire, but the level of agreement also depends to some extent on the instrument used.

In assessing the outcome of more impaired stroke patients or the patients in a more specific OT programme, we believe that the OT needs to give assistance in the use of the self-reported mode of administration used and to make additional assessments, such as observations and interviews.

Computer adapted testing (CAT, building on the Rasch analysis has been introduced in research studies in rehabilitation medicine and might be a future possibility for outcome analyses in clinical settings as well (50). A contemporary alternative mode of administration of questions related to ability might be a computer-based questionnaire, which has shown agreement with face-to-face interviews (12).

A complementary use of the different modes might reduce the time needed for both patients and professionals. Interchangeable use of the methods (semi-structured interviews and self-completed questionnaire) should be considered with caution, if the assessments will be used in the same

patients. The studies described here emphasise how complex and unique each individual ADL situation is. The instability of the ADL assessments is to some extent a method problem, but the studies also indicate the problems that can arise when different raters are involved in the assessments, as is reality in clinical work.

CONCLUSIONS

Methodological studies are important for critically analysing and evaluating clinical methods that assess ADL. A high level of agreement in the ADL assessments could be established between two raters during the same semi-structured interview, but this agreement decreased between different interviews and raters. The result of the studies showed that there were problems with the semi-structured interview as a method for assessing ADL when used by different raters at different occasions. The random subject variation was assumed to interact with systematic rater variations (systematic disagreements in how to interpret the operational definitions). The raters' systematic disagreements were more obvious in the interpretations of the more complex activities such as the social and cognitive items of the FIM™.

By changing the assessment procedure to include a questionnaire, the stability of the ADL assessments could generally be maintained. However, some items that were different in the different ADL instruments still showed poor agreement, indicating a problem with the construct and how well the ADL instruments reflect the phenomena of interest and the interpretation of the concept of ADL independence. Still, the results of the ADL assessments, especially in items with poor agreement, demonstrate a need for a higher level of agreement to reflect true individual changes. The importance should again be stressed of continuing to use the same method in following individual patients to avoid the confusion that can arise in the results if several different methods are used in one patient. A complementary use of a self-reported questionnaire might be an accessible mode of administration in clinical work, but the assessment procedure needs further development to suit each clinical situation, such as acute care.

The ADL instruments of FIM™/IAM and the ADL taxonomy were comparable in terms of discriminating ADL independence, and most ADL dependent persons could be identified regardless of the ADL instrument used. The use of the ADL taxonomy generally showed greater stability in the items when the mode of administration was changed, indicating a more consistent construct for assessing ADL independence. A change in the item difficulty in different settings intensifies the environmental effects in the ADL assessment in a "doing perspective".

Future work

This work can be seen as a part of continuing efforts to develop clinically reliable and feasible ADL assessment procedures. Different raters are a reality in the clinical setting, emphasising the possibility of using a questionnaire to minimise the influence of different raters. The results support a further development of a self-reported questionnaire (on paper or computer-based) being a feasible way in the clinical setting to document ADL ability. More studies are needed to analyse the theoretical constructs of the ADL instruments and to consolidate the concepts concerning how to accurately assess the individual's ADL in/dependence after a stroke. Further studies are also needed to increase the level of agreement between raters and methods to achieve a generally higher stability in all the items included so that it is possible to follow individual changes in reproduced ADL assessments.

SAMMANFATTNING PÅ SVENSKA (summary in Swedish)

Bakgrund

Bedömningar av aktiviteter i dagligt liv, ADL, görs med olika syften vid olika tillfällen inom rehabilitering. Vid upprepade bedömningar i vårdkedjan är det viktigt att veta vad en förändring av aktivitetsnivån beror på. Övergripande syfte var att analysera faktorer som påverkade ADL bedömningen. ADL bedömningen omfattade skattning av personens förmåga att utföra vardagliga aktiviteter självständigt, med eller utan hjälpmedel, eller med hjälp av annan person. Den gjordes för att bedöma om personen efter ett stroke insjuknande behövde assistans för att klara eget boende utanför institution.

Metodik

Olika datainsamlingsmetoder (observation, intervju, post-enkät) och instrument (Functional Independence Measure, FIM™, Instrumental Activity Measure, IAM och ADL-taxonomin) användes. Ett semi-strukturerat tillvägagångssätt användes som grund för FIM- och IAM intervjuerna enligt ett speciellt flödesschema och aktiviteterna bedömdes i sju kategorier. ADL taxonomins delaktiviteter bedömdes i två kategorier ("gör"/"gör inte"). Insamlade data från de ingående aktiviteterna i ADL instrumenten tudelades före den statistiska bearbetningen. Deltagarna i de fyra studierna var en tillgänglighetspopulation av personer som vårdats på en rehabiliteringsavdelning efter ett stroke insjuknande (n=68). Stabiliteten i bedömningarna analyserades avseende skattat ADL utförande utifrån: 1) mellan-bedömar överensstämmelse 2) stabilitet mellan bedömare 3) systematisk oenighet i tolkningen av aktiviteterna 4) överensstämmelse mellan datainsamlingsmetoder 5) överensstämmelse i bedömt ADL o/beroende mellan instrumentens ingående aktiviteter.

Resultat

Resultaten visade relativt god mellan-bedömar överensstämmelse för ett semi-strukturerat intervjuförfarande under samma intervju. Upprepad intervju (FIM™) visade mindre stabila ADL bedömningar. Den hierarkiska ordningen dvs. från lätta till svåra aktiviteter att utföra, ändrades mellan sjukhus- och hemmiljö för personlig vård och instrumentella (boendets) aktiviteter. Detta visar på miljöns inflytande på aktivitet utförandet. Överensstämmelsen mellan två data insamlingsmetoder (post enkät och intervju) var tillfredsställande för de flesta aktiviteter inom personlig vård och för de enklare boende aktiviteterna, förutom några avvikande aktiviteter i varje ADL instrument. I huvudsak identifierades självständighet eller ett beroende av annan person hos samma personer, oavsett vilket ADL instrument som användes.

Konklusion

Flera faktorer framkommer, som är bedömar, instrument och/eller metod relaterade. Studierna visar på komplexiteten i att tolka ett ADL utförande i ett ”gör” perspektiv eftersom det innehåller miljö faktorer som i sig varierar naturligt i ett tidförlopp. Detta kan bidra till olika resultat vid uppföljning och bedömning av ADL

förmåga över tid i olika kliniska sammanhang. Ytterligare studier är nödvändiga för att förtydliga instrumentens struktur och begreppet “skattad självständighet och beroende av annan person” i ett av ”gör” perspektiv.

ACKNOWLEDGEMENTS

I wish to express my warm and sincere gratitude and appreciation to all who have made this work possible and to all fruitful clinical and scientific discussions during the years.

To all my study participants and their proxies for their interest and willingness to share their perceived knowledge and experiences about the stroke impairment.

To my supervisors:

Professor Emeritus Gunnar Grimby, my supervisor who introduced me to and encouraged me in research studies from the very beginning, for his valuable support, guidance and for sharing his deep knowledge of scientific and methodological research.

Professor Katharina Stibrant Sunnerhagen, research leader of the rehabilitation medicine group at the Institute of Neuroscience and Physiology, my supervisor, for her endurance, constructive criticism and encouragement, for always taking time in this long and winding research process and for sharing her extensive knowledge about stroke within the broad field of rehabilitation interventions.

Professor Lena Nordholm, my co-supervisor, for her enthusiasm and fruitful methodological discussions, for valuable support and language revisions.

Lisbeth Claesson, Ph.D. Reg. OT, my co-supervisor, for sharing thoughts and for our fruitful discussions, for support, constructive criticism and always helping and listening to my research questions.

Professor Birgitta Lundgren Lindqvist for introducing me to clinical research studies from the very beginning.

(The following persons are named in alphabetical order)

To all colleagues and friends for valuable comments, support and constructive criticism on the manuscript and fruitful discussions:

Annika Dahlgren, Christopher Lindberg, Åsa Lundgren Nilsson, Ulla Nordenskiöld and Ulla Sonn.

To my co-workers for stimulating discussions and fruitful work:

Eva Andrén, colleagues and friends for sharing many fruitful methodological discussions and for the contribution to this work Ann Björkdahl, Marita Hedberg, Ingela Nordén, Barbro Olsson and Barbro Wikander.

Gunnel Carlsson for sharing thoughts in work in a clinical project with persons after stroke at the very beginning.

To all at the Institute of Neuroscience and Physiology especially:

Christian Blomstrand, Anki Nyberg, Lars Rönnbäck, Kirsten Toftered and Carsten Wikkelsö for all assistance, guidance and Oskar Bergström for valuable technical support.

To my all OT colleagues and friends especially to: Gunilla Forsberg Wärleby, Gunilla Gosman Hedström, Anna-Lisa Thorén Jönsson for sharing thoughts, methodological work and discussions in occupational therapy questions.

To the leaders and all my work-mates at the Occupational Therapy Unit at Sahlgrenska University Hospital during the years, especially thanks to Birgitta Archenholtz, Gunilla Christiansson for support and giving me the possibility of carrying out the research.

Jane Finnstam for encouraging from the start in clinical development work and project studies.

To all my research friends at the Rehabilitation medicine research group, Per Dubbsgatan 14, for sharing ideas and the important informal coffee meetings with fruitful discussions; Anna Ekman for constructive advice and help with statistical analysis; Elisabeth Svensson for valuable comments on the statistical analyses with ROC curves; Valter Sundh for performing part of the Rasch analyses.

To all personal at department of rehabilitation medicine at the Sahlgrenska University Hospital.

To my work-mates and friends at the MS Centrum, the department of neurology at the Sahlgrenska University Hospital for valuable clinical and research discussions.

To Janet Vesterlund for the invaluable help with an extensive and accurate language review of the manuscript.

My dear family Louise and Andreas and friends, for their understanding for my sometimes being distracted and for their patience with there being “a lot of papers lying around”. I am grateful to Andreas for taking the wonderful picture on the cover of the thesis. To Lars Daving for technical advice and support with computer questions.

Financial support

This theses was supported by grants from the Swedish Research Foundation (VR K 2002-27-VX-14318-01A), the Swedish Foundation for Health Care Sciences and Allergy Research (Vårdalstiftelsen), the Council for Research and Development of the Gothenburg Region, the Research Council of the Swedish County Council Federation, the Swedish Stroke Association, the Rune and Ulla Amlow Foundation, the Per-Olof Ahl Foundation, the Greta and Einar Asker Foundation, the Hjalmar Svensson Foundation, the Foundation of the Sahlgrenska University Hospital, the John and Brit Wennerstrom Foundation.

REFERENCES

1. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
2. Archenholtz B, Dellhag B. Validity and reliability of the instrument Performance and Satisfaction in Activities of Daily Living (PS-ADL) and its clinical applicability to adults with rheumatoid arthritis. Scand J Occup Ther. 2008 Mar; 15(1):13-22.
3. Barnes MP, Dobkin BH, Bogousslavsky J. Recovery after stroke. Cambridge: Cambridge University Press; 2005.
4. Bartko JJ, Carpenter WT, Jr. On the methods and theory of reliability. J Nerv Ment Dis. 1976 Nov; 163(5):307-17.
5. Bendz M. The first year of rehabilitation after a stroke - from two perspectives. Scand J Caring Sci. 2003 Sep; 17 (3): 215-22.
6. Borg J. Rehabiliteringsmedicin: teori och praktik. Lund: Studentlitteratur; 2006.
7. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. Bmj. 1992 Jun 6; 304(6840):1491-4.
8. Brennan PF, Hays BJ. The kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. Res Nurs Health. 1992 Apr; 15(2):153-8.
9. Carlsson GE, Moller A, Blomstrand C. A qualitative study of the consequences of 'hidden dysfunctions' one year after a mild stroke in persons <75 years. Disabil Rehabil. 2004 Dec 2;26 (23):1373-80.
10. Carter J, Mant F, Mant J, Wade D, Winner S. Comparison of postal version of the Frenchay Activities Index with interviewer-administered version for use in people with stroke. Clin Rehabil. 1997 May; 11(2):131-8.
11. Chang WC, Slaughter S, Cartwright D, Chan C. Evaluating the FONE FIM: Part I. Construct validity. J Outcome Meas. 1997; 1(3):192-218.
12. Chestnutt IG, Morgan MZ, Hoddell C, Playle R. A comparison of a computer-based questionnaire and personal interviews in determining oral health-related behaviours. Community Dent Oral Epidemiol. 2004 Dec; 32(6):410-7.
13. Christiansen CH, Baum CM, Bass-Haugen J. Occupational therapy : performance, participation, and well-being. 3 ed. Thorofare, NJ: Slack; 2005.

14. Claesson L, Svensson E. Measures of order consistency between paired ordinal data: application to the Functional Independence Measure and Sunnaas index of ADL. J Rehabil Med. 2001 Mar; 33(3):137-44.
15. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological measurement. 1960; 20:37-46.
16. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. Int Disabil Stud. 1988; 10(2):61-3.
17. Dahlin-Ivanoff S, Sonn U, Svensson E. Development of an ADL instrument targeting elderly persons with age-related macular degeneration. Disabil Rehabil. 2001 Jan 20; 23(2):69-79.
18. Dallmeijer AJ, Dekker J, Roorda LD, Knol DL, van Baalen B, de Groot V, et al. Differential item functioning of the Functional Independence Measure in higher performing neurological patients. J Rehabil Med. 2005 Nov; 37(6):346-52.
19. Daving Y, Andrén E, Grimby G. Inter-rater agreement using Instrumental Activity Measure. Scand J Occup Ther. 2000; 7:33-8.
20. DeLisa JA, Gans BM. Rehabilitation medicine: principles and practice. 3. ed. Philadelphia: Lippincott-Raven; 1998.
21. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. Arch Phys Med Rehabil. 1993 May; 74(5):531-6.
22. Domholdt E. Rehabilitation research: principles and applications. 3. ed. St. Louis Mo: Elsevier Saunders; 2005.
23. Donaghy S, Wass PJ. Interrater reliability of the Functional Assessment Measure in a brain injury rehabilitation program. Arch Phys Med Rehabil. 1998 Oct; 79(10):1231- 6.
24. Duncan P, Reker D, Kwon S, Lai SM, Studenski S, Perera S, et al. Measuring stroke impact with the stroke impact scale: telephone versus mail administration in veterans with stroke. Med Care. 2005 May; 43(5):507-15.
25. Duncan PW, Lai SM, Tyler D, Perera S, Reker DM, Studenski S. Evaluation of proxy responses to the Stroke Impact Scale. Stroke. 2002 Nov; 33(11):2593-9.
26. Epstein AM, Hall JA, Tognetti J, Son LH, Conant L, Jr. Using proxies to evaluate quality of life. Can they provide valid information about patients' health status and satisfaction with medical care? Med Care. 1989 Mar; 27(3 Suppl):S91-8.
27. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990; 43(6):543-9.
28. Finlayson M, Mallinson T, Barbosa VM. Activities of daily living (ADL) and instrumental activities of daily living (IADL) items were stable over time in a longitudinal study on aging. J Clin Epidemiol. 2005 Apr; 58(4):338-49.
29. Fleiss JL, Levin BA. Statistical methods for rates and proportions. 2. ed. New York: Wiley; 1981.
30. Fricke JU, C. Worrel, D. Reliability of the Functional Independence Measure with Occupational Therapists. The Australian Occupational Therapy Journal. 1993; 40(1):6-15.
31. Glader EL, Stegmayr B, Johansson L, Hulter-Asberg K, Staaf A, Wester PO. Stroke-and then? Considerable need of assistance two year after stroke according to a large nation-wide study. Lakartidningen. 2001 Oct 10; 98(41):4462-7.
32. Glader EL, Stegmayr B, Asplund K. Poststroke fatigue: a 2-year follow-up study of stroke patients in Sweden. Stroke. 2002 May; 33(5):1327-33.
33. Gosman-Hedstrom G, Svensson E. Parallel reliability of the functional independence measure and the Barthel ADL index. Disabil Rehabil. 2000 Nov 10; 22(16):702-15.
34. Gosman Hedstrom G, Blomstrand C. Evaluation of a 5-level Functional Independence Measure in a longitudinal study of elderly stroke survivors. Disability and Rehabilitation. 2004; 26(7):410-8.
35. Grey N, Kennedy P. The Functional Independence Measure: a comparative study of clinician and self ratings. Paraplegia. 1993 Jul; 31(7):457-61.

36. Grimby G, Andren E, Holmgren E, Wright B, Linacre JM, Sundh V. Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: a study of individuals with cerebral palsy and spina bifida. Arch Phys Med Rehabil. 1996 Nov; 77(11):1109-14.
37. Grimby G, Gudjonsson G, Rodhe M, Sunnerhagen KS, Sundh V, Ostensson ML. The functional independence measure in Sweden: experience for outcome measurement in rehabilitation medicine. Scand J Rehabil Med. 1996 May; 28(2):51-62.
38. Grimby G, Andren E, Daving Y, Wright B. Dependence and perceived difficulty in daily activities in community-living stroke survivors 2 years after stroke: a study of instrumental structures. Stroke. 1998 Sep; 29(9):1843-9.
39. Hamilton BB, Laughlin JA, Fiedler RC, Granger CV. Interrater reliability of the 7- level functional independence measure (FIM). Scand J Rehabil Med. 1994 Sep; 26(3):115-9.
40. Hamilton BB, Granger, CV. Rehabilitation outcomes: analysis and measurement. Baltimore: Brookes; 1987.
41. Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger C. Prediction of rehabilitation outcomes with disability measures. Arch Phys Med Rehabil. 1994 Feb; 75(2):133-43.
42. Heuschmann PU, Kolominsky-Rabas PL, Nolte CH, Hunermund G, Ruf HU, Laumeier I, et al. [The reliability of the german version of the barthel-index and the development of a postal and telephone version for the application on stroke patients.] Comparison of postal version of the Frenchay Activities Index with interviewer- administered version for use in people with stroke. Fortschr Neurol Psychiatr. 2005 Feb May;73(2):74-82.
43. Hochstenbach J, Prigatano G, Mulder T. Patients' and relatives' reports of disturbances 9 months after stroke: subjective changes in physical functioning, cognition, emotion, and behavior. Arch Phys Med Rehabil. 2005 Aug; 86(8):1587- 93.
44. Holm. A simple sequentially rejective multiple test procedure. Scand J Statist. 1979;6:65-70.
45. Hulter Åsberg K. Elderly patients in acute medical wards and home-care. Functional assessment, prediction of outcome, and trial of early activation. Uppsala: University of Uppsala; 1986.
46. Iwarsson S. Environmental influences on the cumulative structure of instrumental ADL: an example in osteoporosis patients in a Swedish rural district. Clin Rehabil. 1998 Jun; 12(3):221-7.
47. Jaworski DM, Kult T, Boynton PR. The Functional Independence Measure: a pilot study comparison of observed and reported ratings. Rehabilitation-Nursing-Research (REHABIL-NURS-RES) 1994 Winter; 3(4): 141-7.
48. Jensen MP, Abresch RT, Carter GT. The reliability and validity of a self-report version of the FIM instrument in persons with neuromuscular disease and chronic pain. Arch Phys Med Rehabil. 2005 Jan; 86(1):116-22.
49. Jette AM. The Functional Status Index: reliability and validity of a self-report functional disability measure. J Rheumatol Suppl. 1987 Aug; 14 Suppl 15:15-21.
50. Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. J Rehabil Med. 2005 Nov; 37(6):339-45.
51. 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. Jama. 1963 Sep 21; 185:914-9.
52. Kazdin AE. Artifact, bias, and complexity of assessment: the ABCs of reliability. J Appl Behav Anal. 1977 Spring; 10(1):141-50.
53. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis. 1985;38(1):27-36.
54. Koran LM. Increasing the reliability of clinical data and judgments. Ann Clin Res. 1976 Apr;8(2):69-73.
55. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar; 33(1):159-74.

56. Law M, Letts L. A critical review of scales of activities of daily living. Am J Occup Ther. 1989 Aug; 43(8):522-8.
57. Law M, Baum C, Dunn W. Measuring Occupational Performance; supporting best practice in Occupational Therapy. Thorafare: Slack incorporated; 2001.
58. Law MC, Baum CM, Dunn W, Law MC. Measuring occupational performance: supporting best practice in occupational therapy. 2. ed. Thorofare, N.J.: SLACK Inc.; 2005.
59. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist. 1969 Autumn; 9(3):179-86.
60. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil. 1994 Feb;75(2):127-32.
61. Lundgren-Nilsson A, Grimby G, Ring H, Tesio L, Lawton G, Slade A, et al. Cross-cultural validity of functional independence measure items in stroke: a study using Rasch analysis. J Rehabil Med. 2005 Jan; 37(1):23-31.
62. Lundgren-Nilsson A, Tennant A, Grimby G, Sunnerhagen KS. Cross-diagnostic validity in a generic instrument: an example from the Functional Independence Measure in Scandinavia. Health Qual Life Outcomes. 2006; 4:55.
63. Mahoney FI, Barthel DW. Functional Evaluation: the Barthel Index. Md State Med J. 1965 Feb; 14:61-5.
64. Mc Dowell I, Newell C. Measuring Health; a guide to rating scales and questionaires. 2nd ed. Oxford: Oxford University Press; 1996.
65. Nachmias D, Frankfort-Nachmias C. Research methods in the social sciences. 2nd ed. New York: St. Martin's Press; 1981.
66. Nilsson AL, Sunnerhagen KS, Grimby G. Scoring alternatives for FIM in neurological disorders applying Rasch analysis. Acta Neurol Scand. 2005 Apr; 111(4):264-73.
67. Nordenskiold U, Grimby G, Hedberg M, Wright B, Linacre JM. The structure of an instrument for assessing the effects of assistive devices and altered working methods in women with rheumatoid arthritis. Arthritis Care Res. 1996 Oct;9(5):358-67.
68. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. Arch Phys Med Rehabil. 1996 Dec;77(12):1226-32.
69. Pedretti LW, Pendleton HM, Schultz-Krohn W. Pedretti's occupational therapy: practice skills for physical dysfunction. 6. ed. St. Louis, Mo.: Mosby Elsevier; 2006.
70. Polit DF, Hungler BP. Nursing research : principles and methods. 4. ed. Philadelphia: Lippincott; 1991.
71. Portney LG, Watkins MP. Foundations of clinical research : applications to practice. 3rd ed. Upper Saddle River, N.J.: Pearson/Prentice Hall; 2009.
72. Rogers JC, Holm MB, Beach S, Schulz R, Cipriani J, Fox A, et al. Concordance of four methods of disability assessment using performance in the home as the criterion method. Arthritis Rheum. 2003 Oct 15; 49(5):640-7.
73. Saxena SK, Ng TP, Koh G, Yong D, Fong NP. Is improvement in impaired cognition and depressive symptoms in post-stroke patients associated with recovery in activities of daily living? Acta Neurol Scand. 2007 May; 115(5):339-46.
74. Schlote A, Kruger J, Topp H, Wallesch CW. [Inter-rater reliability of the Barthel Index, the Activity Index, and the Nottingham Extended Activities of Daily Living: The use of ADL instruments in stroke rehabilitation by medical and non medical personnel]. Rehabilitation (Stuttg). 2004 Apr; 43(2):75-82.
75. Sheikh K. Disability scales: assessment of reliability. Arch Phys Med Rehabil. 1986 Apr; 67(4):245-9.
76. Shrout. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979; 86:420-8.
77. Skidmore ER, Rogers JC, Chandler LS, Holm MB. Dynamic interactions between impairment and activity after stroke: examining the utility of decision analysis methods. Clin Rehabil. 2006 Jun; 20(6):523-35.

78. Smith PM, Illig SB, Fiedler RC, Hamilton BB, Ottenbacher KJ. Intermodal agreement of follow-up telephone functional assessment using the Functional Independence Measure in patients with stroke. Arch Phys Med Rehabil. 1996 May;77(5):431-5.
79. Smith RO. The science of occupational therapy assessment. Occupational Therapy Journal of Research. 1992; 12(1):3-15.
80. Socialstyrelsen. Nationella riktlinjer för strokesjukvård 2005. Stockholm: Socialstyrelsen; 2006.
81. Sonn U. Longitudinal studies of dependence in daily life activities among elderly persons. Scand J Rehabil Med Suppl. 1996; 34:1-35.
82. Sonn U, Törnquist K, Svensson E. The ADL taxonomy - from individual categorical data to ordinal categorical data. Scandinavian Journal of Occupational Therapy. 1999;6:11-20.
83. Stineman MG, Shea JA, Jette A, Tassoni CJ, Ottenbacher KJ, Fiedler R, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. Arch Phys Med Rehabil. 1996 Nov; 77(11):1101-8.
84. Svensson E. Analysis of systematic and random differences between paired ordinal categorical data. Stockholm: Almqvist & Wiksell International; 1993.
85. Svensson E, Holm S. Separation of systematic and random differences in ordinal rating scales. Stat Med. 1994 Dec 15-30;13(23-24):2437-53.
86. Svensson E, Starmark JE, Ekholm S, von Essen C, Johansson A. Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. Neurol Res. 1996 Dec; 18(6):487-94.
87. Svensson E. Ordinal invariant measures for individual and group changes in ordered categorical data. Stat Med. 1998 Dec 30; 17(24):2923-36.
88. Tesio L. Functional assessment in rehabilitative medicine: principles and methods. Eura Medicophys. 2007 Oct 23.
89. Törnquist K, Sonn U. Towards an ADL taxonomy for occupational therapists. Scandinavian Journal of Occupational Therapy. 1994; 1:69-76.
90. Törnquist K, Sonn U, Förbundet Sveriges arbetsterapeuter. ADL-taxonomi : en bedömning av aktivitetsförmåga. Nacka: Förbundet Sveriges arbetsterapeuter; 2001.
91. UDS. Functional Independence Measure, FIMsm, vägledning. 5.0.1996 ed. Gothenburg; 1996.
92. Unsworth C. The concept of function. British Journal of Occupational Therapy. 1993; 56(8):287-92.
93. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford Univ. Press; 1992.
94. van de Port IG, Kwakkel G, Schepers VP, Heinemans CT, Lindeman E. Is fatigue an independent factor associated with activities of daily living, instrumental activities of daily living and health-related quality of life in chronic stroke? Cerebrovasc Dis. 2007; 23(1):40-5.
95. World Health Organization. International classification of impairments, disabilities, and handicaps: a manual of classification relating to the consequences of disease : published in accordance with resolution WHA29.35 of the Twenty-ninth World Health Assembly, May 1976. Geneva: WHO; 1980.
96. World Health Organization. International classification of functioning, disability and health: ICF. Geneva: World Health Organization; 2001.
97. Wressle E, Eeg-Olofsson AM, Marcusson J, Henriksson C. Improved client participation in the rehabilitation process using a client-centred goal formulation structure. J Rehabil Med. 2002 Jan; 34(1):5-11.
98. Wright. Best test design: Rasch measurement. 1979.