Measuring the impact of body functions on occupational performance:

Validation of the ADL-focused Occupation-based Neurobehavioral Evaluation (A-ONE)

Guðrún Árnadóttir

Department of Community Medicine and Rehabilitation, Occupational Therapy, Umeå University
901 87 Umeå, Sweden
2010

ISBN: 978-91-7264-931-6
ISSN: 0346-6612
Printed in Sweden by Arkitektkopia, Umeå, 2010

In memory of Sigga and Baldur R. Stefansson from Manitoba, Canada.

Table of Contents

Abstract

Background: Among the instruments commonly used by occupational therapists working in the area of rehabilitation of persons with neurological disorders are evaluations of both occupation, such as activities of daily living (ADL), and body functions. While persons with neurological diagnoses typically have symptoms that represent diminished neurobehavioral functions, the resulting pattern of neurobehavioral impairments affecting ADL performance often differs among diagnostic groups. Usually, neurobehavioral impairments are evaluated in a context that is separate from and not natural for ADL task performance. The A-ONE is a unique instrument that can be used to evaluate both ADL performance (ADL scale) and, in the natural context of the ADL task performance, the underlying neurobehaviors that cause diminished ADL task performance among persons with neurological disorders (Neurobehavioral scale). The scales of the instrument are of ordinal type, and in their existing form, do not have measurement properties. Measurement properties are a requirement of evidence-based and quality assured rehabilitation services.

The overall aim of this doctoral study was to further develop and validate the A-ONE. This included (a) internal validation to explore the potential for converting the ordinal scales of the instrument to interval scales, (b) examination of which of the neurobehavioral items would be most beneficial and clinically useful for constructing a new Neurobehavioral Impact (NBI) scale for evaluating persons with different neurological diagnoses, and (c) exploration of whether persons with right and left cerebrovascular accidents (RCVA, LCVA) differ in mean NBI measures.

Methods: This thesis is comprised of four studies which all contribute in different ways to the validation of the scales of the A-ONE. In the first three studies, Rasch analyses, a widely accepted modern test theory methodology, was used to examine internal validity of the scales and the reliability of the A-ONE measures. In the fourth study, ANCOVA was used to explore between group differences, and Pearson correlation coefficients were used to explore relations between person measures from the different A-ONE scales.

Results: The first study of 209 persons diagnosed with CVA and dementia provided support for converting the ordinal ADL scale to an interval scale that has potential to be used to measure change in ADL performance over

time. The second and third studies, including 206 and 422 persons respectively, indicated that it is possible to construct several unidimensional versions of a new NBI scale from the neurobehavioral items of the instrument, each with different item content and hierarchical item structure. Further, some of these NBI scales could be used across different diagnostic groups. When exploring differences between 215 persons with RCVA and LCVA on the NBI scale developed for CVA, results of the ANCOVA (with ADL ability as a covariate) indicated that there is no significant difference between groups in their mean NBI measures, despite known differences in patterns of neurobehavioral impairments.

Conclusions: The results of this thesis indicate that the A-ONE, although developed by traditional psychometric methods for the purpose of providing useful information for intervention planning, now also has the potential to be used to measure change and compare diagnostic groups. This additional feature will likely enhance both clinical and research potential of the instrument. In order to make the results of the study accessible for clinicians, conversion tables need to be developed.

Key words: Activities of daily living, Occupational therapy, Rasch measurement, Stroke, Dementia, Evaluation

Abbreviations

A-ONE	ADL-focused Occupation-based Neurobehavioral Evaluation
ADL	Activities of daily living
AERA	American Educational Research Association
AHA	American Heart Association
AMPS	Assessment of Motor and Process Skills
ANCOVA	Analysis of covariance
AOTA	American Occupational Therapy Association
APA	American Psychological Association
ASA	American Stroke Association
Bi	Person ability measure
CNS	Central nervous system
CT	Computerized tomography
CVA	Cerebral vascular accident
CMEEG	Computerized mapping of electroencephalography
CTT	Classical test theory
DAT	Dementia Alzheimer type
Di	Item difficulty calibration
DIF	Differential item functioning
DU	Dementia unspecified
FI	Functional independence
FIM	Functional Independence Measure
ICC	Interclass correlation coefficient
LCVA	Left cerebrovascular accident
LSH	Landspítali University Hospital

Abbreviations (continued)

M	Mean
MnSq	Mean square
MTT	Modern test theory
NB	Neurobehavioral scale
NBI	Neurobehavioral Impact scale
NBPIS	Neurobehavioral Pervasive Impairment Subscale
NBSIS	Neurobehavioral Specific Impairment Subscale
NCME	National Council on Measurement in Education
OT	Occupational therapy
OTIPM	Occupational Therapy Intervention Process Model
PCA	Principal components analysis
RCVA	Right cerebrovascular accident
SE	Standard error

Original Papers

This thesis is based on the following papers:

I Árnadóttir, G. & Fisher, A. G. (2008). Rasch analysis of the ADL scale of the A-ONE. *American Journal of Occupational Therapy*, *62*, 51–60.

II Árnadóttir, G., Fisher, A. G., & Löfgren, B. (2009). Dimensionality of nonmotor neurobehavioral impairments when observed in the natural contexts of ADL task performance. *Neurorehabilitation and Neural Repair, 23*, 579–586.

III Árnadóttir, G., Löfgren, B., & Fisher, A. G. Neurobehavioral functions evaluated in naturalistic contexts: Rasch analysis of the A-ONE Neurobehavioral Impact scale. *Manuscript submitted for publication*.

IV Árnadóttir, G., Löfgren, B., & Fisher, A. G. Difference in impact of neurobehavioral dysfunction on ADL performance between persons with right and left hemispheric stroke. *Manuscript submitted for publication.*

Rationale

I am sitting at a table preparing to have dinner. The person next to me takes an egg from a plate and "occupies" herself by eating the egg without taking the shell off. The person obviously needs assistance to eat effectively, but why, what can I do about it, and how can I evaluate if what I have done altered her performance? Moreover, how can I convince others that there has been a measurable change in her ADL (activities of daily living) task performance after my intervention, so that they will pay me for the intervention? These are some of the most critical issues occupational therapists working with persons with neurological disorders are confronted with.

The ADL-focused Occupation-based Neurobehavioral Evaluation (A-ONE), a standardized criterion-based instrument developed by traditional psychometric methods, was first published in 1990 as an aid for occupational therapists in evaluating persons with neurological disorders. The A-ONE is based on the idea that the occupational therapist is able to use two scales within one instrument to not only identify the person's level of ADL assistance (ADL ability), but also the nature of the underlying neurobehavioral impairments that interfere with the person's ADL task performance. *ADL* are defined in this thesis as self-care tasks (grooming, dressing, and eating), mobility (transfers and getting around inside the home), and functional communication; and *neurobehavior* is defined as any behavior reflecting neurological function.

More specifically, the A-ONE is unique because, in addition to being designed to enable an occupational therapist to determine the level of assistance needed for ADL performance (as can be done using most ADL evaluations), the A-ONE can also be used to evaluate the underlying reason for the lack of independence (Árnadóttir, 1990, 1999, 2004a). That is, while observing and evaluating level of assistance needed for ADL task performances by use of an ADL scale, the occupational therapist can simultaneously observe and evaluate the extent to which neurobehavioral impairments impact the ADL task performance by use of a neurobehavioral scale. This is done by detection of errors in occupational performance that are subsequently classified, using clinical reasoning, by type of

[1] Originally termed Árnadóttir OT-ADL Neurobehavioral Evaluation, where OT refers to occupational therapy

neurobehavioral impairment that impacts task performance in the natural context of the ADL task performances.

As I noted earlier, the A-ONE has been composed of two scales, the Functional Independence (FI) scale and the Neurobehavioral (NB) scale. The FI scale includes 22 5-category rating scale items that are representative of five domains: dressing, grooming and hygiene, transfers and mobility, feeding, and communication. As the FI scale is commonly called the ADL scale, I will refer to it as the ADL scale in the remainder of this thesis. The NB scale includes two subscales, the Neurobehavioral Specific Impairment subscale (NBSIS) comprised of 46 5-category rating scale items, and the Neurobehavioral Pervasive Impairment subscale (NBPIS) comprised of 31 dichotomous items.

Although the A-ONE was originally developed for research use, it later became a clinically practical tool that has been helpful in the process of setting occupational therapy goals and selecting intervention methods. The main reason for the clinical popularity of the A-ONE among occupational therapists is, without doubt, the fact that the occupational therapist becomes able to evaluate underlying neurological body functions in the naturalistic context of ADL task performance, as opposed to the conventional evaluation context where deficit-specific impairments are evaluated separately, and outside the naturalistic context in which they can be observed to impact daily life task performance. The importance of evaluation of neurobehavioral impairments in natural contexts has gained increased support over the last few decades because of the growing awareness that the results of deficit-specific evaluations have limited relationships with daily life task performance in natural contexts (Chaytor & Schmitter-Edgecombe, 2003).

With a changing emphasis in rehabilitation, where financing of services has called for measurement of outcomes, has come the pressure to convert instruments based on ordinal data into measures. Like many tools commonly used in rehabilitation, the A-ONE is not a measure. Rather, it can be used to describe change in performance in a standardized way but not measure the extent of the changes. Thus, there is a need to explore possibilities for converting the A-ONE scales to measures. More specifically, a major reason for implementing this study was that I wanted to determine if (a) the items from the five domains of the ADL scale of the A-ONE and (b) the items on the two Neurobehavioral subscales could be combined and shown to work together to define single unidimensional constructs, one for ADL ability and one that might reflect neurobehavioral impact on ADL task performance. I felt that if this could be done, it would provide occupational therapists with the potential to monitor change based on measurement, and,

in turn, develop an evidence-base as well as quality control of services. The realization of such possibilities would only add to the instrument's already established usefulness. Thus, by building measurement properties into the A-ONE, we could, for example, measure whether the eating performance of the person mentioned in the case sample above improved or got worse. Finally, even when an instrument has sound psychometric qualities, it can be clinically useless. Thus, I felt it important to consider the clinical usefulness of any measures that might be able to be developed when interpreting the results of my research related to improving the psychometric properties of the A-ONE for practice.

Introduction

The person sitting next to me at the dinner table (mentioned in the case sample in the Rationale) now takes a spoon and "occupies" herself by reaching with a spoon to take milk from a glass, instead of grasping the glass, lifting it to her mouth, and taking a sip of milk. This person needs occupational therapy!

Occupational therapy is defined as "the art and science of helping people do the day-to-day activities that are important and meaningful to their health and well-being through engagement in valued occupations" (Crepeau, Cohn, & Boyt Schell, 2003, p. 28). The word *occupation,* as it will be used in this thesis, is defined as engagement in doing (Fisher, 2009); and *occupational performance* is defined as accomplishment of selected activity resulting from the dynamic transaction between the person, context, and task (American Occupational Therapy Association [AOTA], 2008). In this study, the "doing" becomes a process where the person is engaged in a series of goal-directed actions performed over time. Occupation, these chains of goal-directed actions, are used to enable us to occupy space, time, or roles, and they are always observed in the context of daily task performances (Fisher, 2009).

The evaluation of and provision of intervention in relation to performance of daily life tasks are among the most common rehabilitation services provided by occupational therapists for persons with neurological disorders (Geyh, Kurt et al., 2004; Geyh, Cieza et al., 2004; Gillen, 2006; Steultjens et al., 2003). If our profession's focus is occupation, the evaluations used by occupational therapists should focus on evaluating occupation, that is, they should be occupation-based. Thus, we must evaluate people in the context of occupational performance (Fisher, 2009).

The A-ONE, described briefly in the Rationale, is occupation-based, and is used to generate information that not only is useful for the process of occupational therapy (evaluating, setting goals, determining type of intervention, and reevaluating the results), but it also can be used to provide the occupational therapist with information that can be shared with professionals from other rehabilitation disciplines. The ADL task performances (i.e., occupations) observed and evaluated (dressing, grooming and hygiene, transferring and mobility, feeding, and communicating) by using the A-ONE are relevant to virtually all people regardless of which life roles they may have. That is, ADL task performances

(e.g., eating) are important to everyone (or nearly everyone), and are meaningful and valued as such. Eating is not a choice — either the person needs to perform the task independently or with assistance — the only choice the person may have is whether or not he or she will have assistance. With further psychometric development, the A-ONE could play role in measuring change, and providing evidence regarding whether occupational therapists are actually doing the "right things right" in their practice.

In the remainder of this section, I will present the background of my thesis. More specifically, I will first discuss the types of neurological disorders most commonly seen in occupational therapy and the types of neurobehavioral problems they demonstrate. Then I will discuss different types of models used in occupational therapy, with an emphasis on process models, and how the A-ONE fits into the occupational therapy process. Subsequently I will review both traditional and modern methods used for instrument development and validation, as well as instruments used in neurological rehabilitation, including both ADL scales and evaluations of neurological body functions. I will conclude with a brief review of the development of the A-ONE placed in the context of Yerxa's (1994) model of an integrated profession.

Neurological disorders seen in rehabilitation

Occupational therapy services are an integral part of the health care system for persons with neurological disorders. This includes persons diagnosed with cerebrovascular accidents (CVA) and dementia, the two most common types of diseases resulting in neurological impairments and disability (American Heart Association [AHA], 2009; National Institute of Neurological Disorders & Stroke, 2010; Rijken & Dekker, 1998).

CVA is a disease resulting in disruption of blood and oxygen supply to brain cells. Impaired function resulting from CVA can be related to dysfunction of areas supplied by the major cerebral arteries in the two hemispheres , or arteries supplying subcortical structures. Impairments may restrict motor, sensory and visual functions, perception, cognition, language, and emotional functions (American Stroke Association [ASA], 2009a, 2009b; Bartels, 2004).

Estimates from the World Health Organization (AHA, 2009) indicate that 15 million people suffer stroke each year, and 5 million of those are left permanently disabled. In the last 2 decades, the actual number of stroke deaths has declined, partially due to improved acute stroke care (AHA, 2009; Langhorne, Williams, Gilchrist, & Howie, 1993) despite of increased

stroke incidence (Medin, Nordlund, & Ekberg, 2004). Thus, the number of persons needing rehabilitation is expected to increase.

Stroke-related medical and disability costs, including hospital costs, lost wages, and decreased productivity, is huge (ASA, 2009c; Claesson, Lindén, Skoog, & Blomstrand, 2005; Heart and Stroke Foundation of Canada, 2009). The majority of the cost in Britain (80%) was related to inpatient hospital care and residential care (British Heart Foundation Statistics, 2009). In Sweden, rehabilitation accounts for 17% the cost of stroke services (Sundberg, Bagust, & Terént, 2003).

Occupational therapists work with persons diagnosed with CVA in acute care facilities, rehabilitation centers, and through community and outpatient services as well as in long-term care facilities. Their main role is evaluation and intervention to diminish the effects on stroke on daily life task performance, and reevaluation to evaluate outcomes (Gillen, 2006; Rijken & Dekker, 1998; Schultz-Krohn & Pendleton, 2006; Steultjens et al., 2003).

Dementia is a collective term used to describe degenerative cognitive brain disorders resulting from different syndromes or conditions. Dementia of Alzheimer's type (DAT) is the most common form of dementia, accounting for 50–70% of cases (Alzheimer Society of Canada, 2009). Six million North Americans are reported to have DAT and this number is expected to double within the next 25 years (Alzheimer's Association, 2009; Alzheimer Society of Canada, 2009). All types of dementia are progressive, resulting in neural cell deterioration and cell death (Alzheimer's Association, 2009). Consequently, neurobehavioral functions become restricted, affecting the person's ability to engage in and perform daily life tasks. DAT and related dementias are reported to be the third most expensive disease to treat in the United States. Limitations in ADL performance have been found to be an important predictor of cost, caregiving services and cost increasing with the severity of the disease (Taylor, Schenkman, Zhou, & Sloan, 2001).

Occupational therapists work with persons diagnosed with dementia in geriatric hospital wards, through outpatient and community services, as well as in long term care facilities. As with persons with CVA, their role is evaluation, intervention, and reevaluation of effectiveness of intervention and/or monitoring for signs of deterioration requiring an updated intervention program.

Referrals to occupational therapy for persons who have had CVAs or suffer from dementia are usually made when the resulting impairments are suspected to affect daily life task performance (Árnadóttir, 2004a). When neurobehavioral impairments occur and limit daily life task performance, the pattern of impairments related to the different diagnostic groups and

even diagnostic subgroups can be quite different. Some of these differences are summarized in the next section.

Patterns of impairments in CVA and dementia

Dysfunction of neurobehavioral functions as a result of CVA may interfere with primary ADL. Impairments appear immediately after the CVA and may diminish over time, although residual impairments are common. Persons diagnosed with CVA have different patterns of impairments depending on the cortical side and/or brain areas affected. Thus, persons diagnosed with RCVA more frequently have visuospatial impairments, unilateral neglect, and motor and sensory problems affecting the left body side. Persons diagnosed with LCVA, on the other hand, more often are reported to have aphasia, apraxia, and unilateral sensory and motor problems affecting the right body side (ASA, 2009b; Bartels, 2004; Caplan, 1993).

Impairments that can be related to the diagnosis of dementia usually develop more slowly and do not limit ADL task performance in the earliest stages of the disease. Tasks classified as instrumental ADL and leisure tasks are impacted first (Gauthier & Gauthier, 1990; Taylor et al., 2001). The progressive decline in body functions and the resulting impairments are frequently related to different stages of the disease. The patterns of impairments detected in dementia can also be related to different subtypes. The earliest noticed impairments in Alzheimer disease relate to organization and sequencing of task steps, memory functions, language functions, and emotional signs. As the disease progresses, memory and language impairments increase, apraxia and visuospatial problems become apparent, and perseverative errors begin to emerge. Judgment and insight into one's own performance limitations also become affected. In the final stages, all neurobehavioral functions, including motor functions may be impaired (Alzheimer's Association, 2009; Árnadóttir, 1990).

In summary, it is readily evident that the number of persons suffering from CVA and dementia is enormous. This number is expected to increase tremendously in unchanged conditions within the next few years, and the involvement of occupational therapists in rehabilitation services for these individuals will be required on a larger scale. Persons with CVA or dementia are the ones occupational therapists most often need to (a) evaluate at the beginning of the occupational therapy process, and (b) reevaluate later on in the process to record changes and determine effectiveness of services. Thus, sound instruments that can be used to detect both occupational performance problems and measure change, be it at the level of occupational performance

or impact of impaired body functions on occupational performance, are needed to meet this demand.

Conceptual and process models in occupational therapy

Occupational therapists use different types of models or frameworks for professional reasoning when they work with persons diagnosed with CVA or dementia. Among them are conceptual and process-driven practice models that mold our thinking and guide our practice. A *conceptual model* is defined by Kielhofner (2009) as an interrelated "body of theory, research and practice resources" (p. 13) originally challenged by practice. A conceptual model, thus, includes the knowledge and theoretical principles that enable occupational therapists to understand the occupation-related problems people are having and how to work with them to overcome their problems.

Process models, on the other hand, guide the delivery of occupational therapy services including evaluation, intervention, and monitoring outcomes (AOTA, 2008; Kielhofner & Forsyth, 2008). Hagedorn (2001) points out that when theory is used to drive practice, a conceptual framework is selected before the nature of the problem has been determined. Thus, framework selection will affect and limit the instruments used and actions taken for intervention. In contrast, when a process model is used to drive practice, the occupational therapy process is used to determine the nature of the problem and to decide how to deal with it through intervention. As a result, choices are made between available conceptual models and approaches for evaluation, as well as planning and implementing most applicable interventions as the process progresses (Hagedorn, 2001).

Several process models that have been published within the discipline of occupational therapy in the last 2 decades. These include the Canadian Practice Process Framework (Davis, Craik, & Polatajko, 2007), Model of Human Occupation (Kielhofner, 2009), Occupational Adaptation (Shultz & Schkade, 2003), Occupational Functioning Model (Trombly Latham, 2008a), and Occupational Therapy Intervention Process Model (OTIPM) (Fisher, 1998, 2009). The OTIPM is different from most of the other process models in two important aspects. First, this model specifies that the occupational therapist must observe the person's performance of naturalistic daily life tasks and implement performance analyses. *Performance analysis* refers to evaluation of the quality of a person's occupational performance as observed by a therapist (Fisher, 2009).

Second, after defining and describing the quality of the goal-directed actions of a particular daily life task performance, and before selecting one or more conceptual models for intervention, the therapist proceeds to interpreting the cause of the person's performance problems (i.e., impaired body functions, person factors, environmental factors). Subsequently, based on goals established during the evaluation phase, the occupational therapist selects a model for planning and implementing intervention. These aspects of the OTIPM also differ from most of the other process-oriented models where choice of a conceptual model for intervention (theory-driven as opposed to process-driven reasoning) takes place before evaluation (Fisher, 2009; Hagedorn, 2001).

The OTIPM also specifies that the intervention should be occupation-based and could be based on a variety of different conceptual models including compensation, acquisition, restoration, and/or education. The final aspect of the process defined in the OTIPM, as in most other process models, is to reevaluate occupational performance, and thereby, provide a basis for evaluation of program effectiveness and/or generating evidence. Thus, the OTIPM is a top-down approach, which includes performance analysis, task analysis, and activity analysis at different phases of the process (Fisher, 1998, 2009).

Placing the A-ONE in the occupational therapy process

The A-ONE administration is compatible with the OTIPM as it is administered in a top-down manner, starting with performance analysis using nonstandardized terminology to describe errors in ADL task performance. This information subsequently provides the basis for task analyses, where the errors are related to operational definitions of neurological impairments hypothesized to be the cause of the person's problems with ADL task performance.

The very fact that the OTIPM requires observation of occupational performance is important, in particular when working with persons with neurological problems. This is, for example, because persons with neurological diagnoses have been reported to sometimes lack insight into their own problems (Burgess, Alderman, Evans, Emslie, & Wilson, 1998; Chaytor, Schmitter-Edgecombe, & Burr, 2006). Direct observation of performance by a professional further eliminates bias in caregivers' judgments (Bouwens et al., 2008; Doble, Fisk, & Rockwood, 1999).

Information based on results from the A-ONE is useful in the occupational therapy process regardless of which of the four intervention

models described by Fisher (2009) is chosen. That is, information on level of assistance needed, quality of performance (errors), and impairments impacting performance (cause) can be addressed when planning the intervention program regardless of whether the program includes adaptive, acquisitional, and/or restorative occupation, or involves the use of an occupation-based education program (e.g., an educational seminar or workshop for families or caregivers of persons with neurological impairments). Thus, the A-ONE fits within the OTIPM process in terms of providing information for goal-setting and providing useful information for intervention. But, in its current form, the A-ONE is not suitable for evaluating change clinically, which is the final step in the OTIPM. This limitation was a critical factor underlying the need for the research presented in this thesis.

Clinical reasoning and the A-ONE

When applying the A-ONE in practice to evaluate ADL task performance, and subsequently neurobehavioral impairments that limit ADL task performance, the occupational therapist applies different types of clinical reasoning (Árnadóttir, 1999, 2004a). These include, for example, *procedural reasoning* (Mattingly & Fleming, 1994), referring to hypothesis formation following interpretation of cues about the nature of problems that interfere with occupational performance. In other words, observed errors in ADL task performance are used to help identify the cause of the activity limitation.

Measurement theories and definitions of terms

Development of instruments used in rehabilitation, including occupational therapy has, to date, mainly been based on two different measurement theories. These are the classical test theory (CTT) that has been used in the past for development of most instruments used within rehabilitation services, and modern test theory (MTT). Both CTT and MTT are used for developing instruments designed to define and assess a *latent variable* (McAllister, 2008) — a construct representing an unobservable characteristic of the people tested (Wolf & Smith, 2007a). However, there is a fundamental difference in the approaches used for assessing the latent variable when CTT as opposed to MTT is applied for that purpose (McAllister, 2008), and the psychometric qualities of instruments developed by CTT have been criticized when it comes to their use for measurement.

This criticism is based primarily on issues related to linear versus ordinal data, which I will discuss in more detail below.

Measurement versus ordinal scores

Measurement is defined as "the location of objects along a single dimension on the basis of observations which add together" (Bond & Fox, 2007, p. 312). Although the CTT is based on classifying different levels of the latent variable, based on their qualities, using numerical ordinal codes, it is important to realize that only interval and ratio scales have measurement properties (Bond & Fox, 2007; W. P. Fisher, 1993; Merbitz, Morris, & Grip, 1989; Wolfe & Smith, 2007a; Wright & Linacre, 1989). In other words, it is only possible to apply mathematical manipulations (e.g., adding up of raw item scores and calculating the difference between two such sums), if interval or ratio scales have been used for the evaluation. Thus, only such scales provide the basis for the possibility of measuring change.

For comparison, *ordinal scores* only provide information regarding whether there is a difference in the order of the numerical values assigned for the purpose of describing a condition (e.g., 4 = no assistance, 3 = verbal assistance, 2 = physical assistance, 1 = total assistance). That is, ordinal scores only provide information regarding whether one value is more or less (e.g., better or worse) than another, not how much more or less. Ordinal scales, therefore, are useful for describing a condition, but not for measuring it.

In rehabilitation, increased emphasis is being placed on the use of scales that have measurement properties (Tesio, 2003), both for clinical and research use. Consequently, an increased number of scales developed using the MTT are now being used in the field (Bond & Fox, 2007; Lim, Rodger, & Brown, 2009; Tesio, 2003). MTT has also been used to re-validate instruments originally developed using CTT methods in hopes of improving or expanding their psychometric qualities and developing their measurement potential (Bond & Fox, 2007; Wright & Linacre, 1989).

Reliability and validity

Whether CTT or MTT methods are used, evaluation of an instrument must include examination of aspects of reliability and validity. *Reliability* refers to the consistency of total scores when a testing procedure is repeated on individuals or groups (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999; Haertel, 2006). The reliability of the total scores for groups is frequently based on internal consistency

reliability calculated using Cronbach's coefficient alpha (DeVellis, 2006). Other familiar methods for evaluating the reliability of total scores are parallel forms (Haertel, 2006) and test-retest reliability (Anastasi & Urbina, 1997). Finally, interrater reliability pertains to the consistency of scores when people are rated by different raters (Golden, Sawicki, & Franzen, 1990).

The concept of validity has evolved such that, over past 2 decades, the definition of validity is now used to refer to the unitary concept of construct validity (vs. the earlier three types of validity: content, criterion-related, and construct validity). *Validity* is defined as the degree to which accumulated evidence and theory support the interpretation of test scores for the proposed purpose (AERA et al., 1999). Further, integrated validity evidence should be used to examine different aspects of validity in view of the following five types of evidence, based on: (a) test content, (b) response pattern (process), (c) internal structure, (d) relations to other variables, and (e) consequences of testing. Thus, validation is perceived as an investigatory process relying on evidence to support the instrument's intended use and interpretation of measures (Wolfe & Smith, 2007b).

Classical test theory

Many instruments used in health research have been developed using CTT (DeVellis, 2006). The use of CTT for instrument development in occupational therapy can be related to a four-step process (Benson & Clark, 1982). These steps are planning, construction, quantitative evaluation, and further validation. It is during the latter two stages that ordinal raw scores from a sample are evaluated through research studies of reliability and validity.

A requirement of a scale developed within CTT is that it includes multiple items that are substantially correlated with each other, and thus the items comprising the scale can be viewed as being unidimensional. Factor analysis, a method used to detect factors that are based on items that correlate with one another in a meaningful way (Linacre, 1998), is the primary statistical method used to assess dimensionality within CTT (DeVellis, 2006).

An important assumption underlying the statistical methods used in CTT is that the data used are interval data, not ordinal. However, most statistical analyses of instruments used within rehabilitation are based on ordinal scales (Fänge, Lanke, & Iwarsson, 2004; Tesio, Simone, & Bernardinello, 2007). While there are different views regarding how to deal with ordinal data statistically, many support the view that ordinal data should not be dealt with using parametric methods, despite numerous examples of such

use in published literature (Davies & Gavin, 1998; Merbitz et al., 1989; Smith, 2001).

Classical test theory applied to the A-ONE

The A-ONE was developed using CTT as were most instruments at the time of its development (see Table 1). The purpose of the A-ONE was to provide information useful for decision-making in relation to setting goals and choice of intervention methods, and the results used to "describe" change in scores, but not to "measure" differences. I recognized that summing up of ordinal scores was not valid, and therefore, the use of summed total scores was discouraged when using the A-ONE.

Critique of classical psychometric methods

One of the main concerns related to scales used in rehabilitation and developed by CTT is, as stated above, that many of these scales are ordinal scales without measurement properties; they are not interval scales. Thus, these scales cannot be used to measure outcomes. Further, many of the statistical methods used in CTT require interval scales. Using inappropriate mathematical manipulations (e.g., adding up ordinal scores, as if they had equal intervals, to form a total score) has been criticized (Merbitz et al., 1989) as the results lack meaning and can result in serious misinterpretation of the results.

Other concerns of the CTT include that the method is sample-dependent, as item difficulties are based only on the sample's performance (Bond & Fox, 2007; DeVellis, 2006; McAllister, 2008). Therefore, group comparisons can be problematic. Yet, the use of a single scale to measure performance of groups with different diagnoses may be desirable (Tennant et al., 2004; Tesio, 2003) for reporting the rehabilitation outcomes in comparable terms. For example, such possibilities can be of interest in the context of evaluating overall rehabilitation outcomes or comparing intervention programs between groups. But the ordinal scales based on CTT can only be used for documenting and describing a condition or therapeutic effectiveness in qualitative terms, cannot be used for measuring change or making comparisons between groups. Therefore, although ordinal data are being gathered, such data needs to be transformed by use of MTT if it is to be used for measurement.

Table 1 Evidence for Reliability and Validity of the A-ONE based on Classical Test Theory

Study	Type of evidence	Results
CTT phase II: Construction Árnadóttir, 1990	**Evidence based on test content:** Content validation	Literature review Expert opinion regarding content of domains
CTT phase III: Quantitative evaluation Árnadóttir, 1990	**Reliability:** Interrater reliability	FI scale: - Average kappa coefficient *(κ)* = 0.83 NB scale: - Average kappa coefficient *(κ)* = 0.85
CTT phase III: Quantitative evaluation Árnadóttir, 2005	**Reliability:** Interrater reliability	FI scale: - *ICC* = 0.98 - Kendall's *τ* = 0.92 - *κw* = 0.90 NBI specific scale: - *ICC* = 0.93 - *κw* = 0.74
CTT phase III: Quantitative evaluation Árnadóttir, 1990	**Evidence based on internal structure:** Inter-item correlations	- FI within domains: $r = 0.3 – 0.9$ - FI across domains: $r = 0.1 – 0.8$ - Percentage of frequencies of significant item correlations across scales (ADL/NB): 75% ($p \leq 0.05$)
CTT phase III: Quantitative evaluation Steultjens, 1998	**Evidence based on internal structure:** Construct validation	- ADL domains: High internal consistency, Cronbach's alpha coefficients = 0.82 – 0.93 - Communication not related to ADL domains
CTT phase IV: Validation Árnadóttir, 1990	**Evidence based on relation to other variables:** Construct validation	Exploratory factor-analysis: - FI scale: 3 factors - NBSI subscale: 2 factors Internal consistency - FI scale α range = 0.75 – 0.79 - NBSIS α range = 0.69 – 0.75 - NBPIS scale α range = 0.59 – 0.63
CTT phase IV: Validation Steultjens, 1998	**Evidence based on relation to other variables:** Concurrent validation	- Correlations of A-ONE FI scale and Barthel Index, $r = 0.85$ - Correlations of A-ONE NB scores and MMSE, $r = 0.70$
CTT phase IV: Validation Gardarsdóttir & Kaplan, 2002	**Evidence based on relation to other variables:** Concurrent and construct validation:	- Difference in ADL: 1/20 item ($p \leq .05$). - Difference in NBSIS ($p \leq .05$): 13 (unilateral body and

Table 1 (continued)

Study	Type of evidence	Results
Gardarsdóttir & Kaplan, 2002 (continued)	- Explore difference in performance of persons with RCVA and LCVA - Explore which NBI items interfere most frequently with ADL	spatial neglect, motor and ideational apraxia, organization and sequencing) - Most frequently detected items: Organization and sequencing, Spatial relations impairment, Unilateral body neglect, Wernicke's aphasia, Broca's aphasia
CTT phase IV: Validation Nuwer et al., 1994	Evidence based on relation to other variables: Concurrent and construct validation: - Explore association of therapists' hypothesis about lesion location based on clinical observations and results of technological evaluation methods	- A-ONE to CT scans, $\kappa = 0.75$ - A-ONE to CMEEG, $\kappa = 0.63$ - CT to CMEEG = $\kappa = 0.53$

Modern test theory and Rasch analysis

Rasch measurement methods are based in one modern test theory that is intended to prevent the problems inherent to CTT by transforming ordinal data into equal interval measures expressed in linear log-odds probability units (logits) (Rasch, 1960/1980; Wright & Linacre, 1989). That is, what has been referred to as MTT includes both item response theory and Rasch measurement. In rehabilitation, it has become customary to use Rasch

Rasch measurement and analysis methods are a family of statistical models used in the development of new assessment methods and in the evaluation of existing instruments developed by CTT. Rasch analysis methods are also commonly used to evaluate various forms of construct validity (Bond & Fox, 2007; Lim et al., 2009). Rasch analysis procedures, based on the original work of George Rasch (Rasch, 1960/1980), have been described elsewhere in detail (Bond & Fox, 2007; Wright & Masters, 1982; Wright & Stone, 1979).

The choice of appropriate model from the "family" of different Rasch models is based on different scoring models and number of facets (Bond & Fox, 2007). All Rasch models assume that some items included in a test will be more difficult than others. The simple Rasch model is based on two assertions referring to the probabilistic relationship between any item's difficulty (*Di*) and any person's ability (*Bn*). These assertions underlie

unidimensionality. *Unidimensionality* refers to the idea that items included in an instrument must define a single construct that is represented by a hierarchy of items that are arranged from those that are easily performed to those that are hard to perform (Bond & Fox, 2007; Wright & Linacre, 1989). If the hierarchy is unidimensional, equal intervals between items along the scale can be assumed (Smith, 2001).

Unidimensionality can be evaluated by use of Rasch-based goodness-of-fit statistics indicating how well the data fit the Rasch model assumptions. That is, when the items demonstrate statistical goodness of fit to the Rasch model, there is some evidence to support unidimensionality of the scale (Bond & Fox, 2007; Wright & Linacre, 1989). Fit is determined by exploring deviations of each item's and person's residual responses from the expectations of the Rasch model used. Two alternative statistics indicate the degree of fit of an item or a person to the modeled underlying construct, the standardized (*z*) statistic and mean squares (*MnSq*). *MnSq* is the mean of the squared difference between what is observed and expected.

Unidimensionality is also evaluated by using principal components analysis (PCA) of the residuals (Smith, 2000). PCA of Rasch-based residuals is used to examine contrasts between opposite (positive vs. negative) loadings of deviations from the modeled Rasch construct (i.e., Rasch factor) that explains most of the variance (Linacre, 1991-2006). Thus, such analyses are different from traditional factor analyses used in CTT where the concern, as mentioned earlier, is to detect factors based on items that correlate with one another in a meaningful way. Factor analysis cannot be used to construct linear measures (Linacre, 1998).

Finally, unidimensionality can be explored through analysis for differential item functioning (DIF) (Smith, 2000). DIF is often defined as a statistically significant difference in item performance among persons from different groups or subgroups which have the same ability level on the underlying construct measured by the scale (Conrad, Dennis, Bezruczko, Funk, & Riley, 2007). DIF is of most concern when it results from factors irrelevant to the construct being measured. In such situations, DIF can result in unfairness, a situation where one group has an unfair advantage over another (Camilli, 2006; Penfield & Camilli, 2007; Perrone, 2006). DIF can, however, also represent a diagnostic indicator where persons from two groups actually display different patterns (Conrad et al., 2007). Thus, persons from different diagnostic groups (e.g., RCVA, LCVA), matched on the basis of having obtained the same total score on a neurobehavioral evaluation, might have significantly different item performance on diagnosis-specific impairment items (Cella & Chang, 2000; Conrad et al.,

2007). In such cases, item hierarchies will differ between groups, and DIF can be said to be present. However, as long as diagnosis-related DIF remains balanced (i.e., such that a group which obtains higher scores on some items also obtains lower on others), the resulting measure can often be shown to remain fair to members of both groups in terms of measuring the degree to which they manifest the underlying construct. That is, the presence of statistical DIF does not always disrupt the measurement system by producing what has been termed *differential test functioning* (DTF) (Borsboom, 2006; Penfield & Camilli, 2007; Tennant & Pallant, 2007).

Additional evidence for scale validity can be provided by (a) verification of logical hierarchical ordering of the items along the linear scale, based on the item difficulty calibration values; and (b) the targeting, referring to how well the item difficulties are aimed at the performance level of the target population (Wright & Stone, 1979). Further, a high quality measure requires a statistical assessment of the psychometric properties of the rating scale (Linacre, 2002; Tennant, 2004).

Rasch analysis computer programs also generate reliability estimates for both persons and items. More specifically, Rasch measurement models assume that some error will occur as a result of human variability. Thus, the standard error (*SE)* related to estimating the location of both each person and each item is calculated and used as an index of reliability. For people, the *SE* becomes important for sensitivity of the estimated measures when used for evaluating change.

Information about reliability generated by Rasch analyses is further revealed by two indexes in the form of a reliability coefficient *(R)* and a separation index *(G)*. The reliability coefficient indicates replicability of person or item placements along the scale. The separation index indicates spread or separation in *SE* units. The separation index for persons indicates how well the items separate the entire sample of people into statistically distinct levels of ability. Similarly, item separation is an index of how well the people separate the items into different levels of difficulty. High separation indicates a scale that covers a wide range of the construct being measured. Thus, the smaller the *SE*, the more likely the generated measures will be reliable and sensitive indices of change (Bond & Fox, 2007).

Another advantage of Rasch measurement methods is that they are both test- and sample-free. Test-free measurement refers to the idea that if the same persons are evaluated using a similar test, one could expect that they would obtain the same hierarchical results. Sample-free (as opposed to classical sample-dependent) measurement refers to the idea that the item difficulty arrangement will not vary significantly between samples, provided

they come from the same population. Thus, if another sample of persons with dementia or CVA were to be evaluated by the A-ONE, one could expect the same hierarchical ordering of the items along the A-ONE scales.

Purpose of evaluation

A crucial prerequisite for choosing an instrument is determining for which purpose it is to be used. Both for gathering information to aid in goal setting and choosing type of intervention, either standardized or nonstandardized instruments may serve the purpose. However, if the purpose is to measure change in performance, standardized instruments and ordinal scales are not enough despite reported validity and reliability (Lim et al., 2009; McAllister, 2008).

In the preceding sections, I have noted that many instruments used in rehabilitation have limited measurement properties, as they were developed by CTT and have ordinal scales; the A-ONE is one of them. Further, such scales can potentially be converted to interval scales and measures by performing Rasch analyses. In the next section, I will turn the discussion to some of the instruments used in rehabilitation and how the A-ONE scales compare to those. This includes both instruments used to evaluate ADL task performance and neurological body functions.

Instruments used in neurological rehabilitation

ADL scales

There are a number of ADL scales in use in rehabilitation (Asher, 2007; Gillen, 2009; Law, Baum, & Dunn, 2005; Neistadt, 2000; Unsworth, 1999), and these scales are used for different purposes such as evaluating ability and level of performance, need for services or intervention, change in performance, prediction of performance, and cost effectiveness. To explore how the A-ONE ADL scale compared to other ADL scales used in rehabilitation, a literature review of instruments best suited for use in occupational therapy to evaluate change in ADL performance of adults with neurological disorders was performed (Árnadóttir, 2008). This review included (a) comprehensive occupational therapy texts and reviews of evaluation methods used in occupational therapy, and (b) Web search including information based on meta-analysis and systematic reviews of outcome studies in occupational therapy. Criteria were set to enable classification and comparison of the obtained information. The criteria included a requirement for observation of ADL task performance, interval scaling, internal validity, and acceptable coefficients for interrater reliability.

Additionally, use of occupational therapy concepts useful for decision-making within occupational therapy, such as ADL, performance skills, or body functions (c.f. AOTA, 2008), was considered to be an asset

Only three out of 24 possible instruments were determined appropriate for further consideration. These instruments included the Assessment of Motor and Process Skills (AMPS), Barthel Index, and Functional Independence Measure (FIM™). Rasch analysis has been performed on all three scales (A. G. Fisher, 1993, 2006a, 2006b; Heinemann, Linacre, Wright, Hamilton, & Granger; 1993; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994; de Morton, Keating, & Davidson, 2008; Nilson, Sunnerhagen, & Grimby, 2005), but for the Barthel and FIM, this information is not readily available for clinical use. The AMPS has available a many-faceted Rasch computer program that trained raters use for analyzing the results from the evaluation (Fisher, 2006b). A review of the literature revealed no published raw-score-to-logit conversion tables for either the FIM and Barthel. Thus, despite the fact that both tools have been subjected to Rasch analysis (Fisher et al., 1994; Heinemann et al., 1993; Linacre et al., 1994; de Morton et al., 2008; Tennant, Geddes, & Chamberlain, 1996), the apparent result is that total raw scores continue to be used in clincial applications of both tools. Further, both instruments have misfitting items (Fisher et al., 1994; Heinemann et al., 1993; Linacre et al., 1994; de Morton, et al., 2008; Tennant et al., 1996). In earlier FIM studies, the social and communication items did not fit on a scale with ADL items (Fisher et al., 1994; Linacre et al., 1994).

Both the FIM and the Barthel Index are generic ADL instruments, used across diagnostic groups. They have commonly been used in neurological rehabilitation outcome research (Geyh, Kurt et al., 2004; Geyh, Cieza et al., 2004; Haigh et al., 2001; Steultjens, Dekker, Bouter, Cardol et al., 2003;).

The AMPS (Fisher, 2006b) is also used across diagnostic groups. Unlike the FIM and Barthel, the AMPS was developed within the discipline of occupational therapy and is used to evaluate quality of observed ADL task performance, not just need for assistance, a construct important for intervention planning in occupational therapy.

When compared to the A-ONE, the AMPS is used only to evaluate ADL, but does so at a more discrete level than does the ADL scale of the A-ONE. The A-ONE, on the other hand, can be used to evaluate both ADL and underlying body functions based on the ADL observation. The two instruments are complementary in that AMPS provides specific information on performance skills (smallest observable units of ADL task performance), and the A-ONE on neurobehaviors limiting ADL task performance. Both

instruments can be easily integrated into practice when the OTIPM is used as the process model. Because of the occupational therapy focus and emphasis on observation of ADL task performance, the AMPS and the A-ONE were determined to be the instruments best suited for use in occupational therapy to measure change in ADL performance of persons with neurological disorders (Árnadóttir, 2008). Thus, there is no doubt in that the A-ONE has place in neurological rehabilitation. That place would only be strengthened if the ADL and NB scales can be converted to linear measures.

Scales for evaluation of neurological impairments

Scales used for evaluation of neurobehavioral impairments can be classified into scales administered in (a) a conventional test context isolated from natural daily life task performance, and (b) a more natural, ecologically-relevant performance context in terms of tasks, tools, and settings. The importance of evaluating neurobehavioral impairments in naturalistic contexts has gained increased support in the literature, within occupational therapy (Gillen, 2009; Neistadt, 2000), as well as neurology and neuropsychology (Bouwens et al., 2008; Semkovska, Bédard, Godbout, Limoge, & Stip, 2004; Schwartz, Mayer, FitzpatricDeSalme, & Montgomery, 1993; Wilson, 2002).

Only a few standardized instruments have been designed to evaluate neurobehavioral impairments in naturalistic contexts. Examples include assessments aimed at evaluating a limited range of impairments such as the Melbourne Low-Vision ADL Index (Haymes, Johnston, & Hayes, 2001), which is a measure of visual dysfunctions only, and the Assessment of Awareness of Disability (Tham, Bernspång, & Fisher, 1999), used to evaluate client's level of insight based on comparing his or her self-reported ADL performance to that actually observed by an examiner. Neither of the above addresses global neurobehavioral functions. The Naturalistic Action Test (Schwartz, Segal, Veramonti, Ferraro, & Buxbaum, 2002) can be used to evaluate everyday action errors associated with executive functions. Similarly, functional sequencing ability while preparing a single meal can be evaluated by the Rabideau Kitchen Evaluation-revised (Neistadt, 1992). In summary, none of the available instruments reviewed, other the A-ONE, can be used to evaluate the wide range of neurobehaviors that can be observed to impact the quality or level of naturalistic ADL task performance.

Ecological validity

Traditionally, conventional and isolated methods have been used to evaluate neurobehavioral impairments (e.g., executive functions, visuospatial functions, motor functions). Such practice places at risk the "ecological validity" of the results (Johnstone & Frank, 1995). To ensure that the neurobehavioral impairment test results can be generalized to natural contexts of ADL task performance, studies examining ecological validity must be completed (correlations of task performance to neurobehavioral impairments) to explore the effect of neurobehavioral impairments on ADL (Cooke, McKenna, Fleming, & Darnell, 2006; Johnston, Findley, DeLuca, & Katz, 1991; Spooner & Pachana, 2006). According to Hammond (1998), the concept of ecological validity was originally introduced to describe the informativeness of cues (i.e., the correlation of a cue and a related variable in an experiment). The term is now commonly used in the literature to refer to the degree to which test performance corresponds to "real world" performance. Ecological validity does not refer to the test itself, but rather the inferences drawn from the test and the utility of those inferences (Chaytor & Scmitter-Edgecombe, 2003).

Two approaches are commonly used in attempt to establish ecological test validity (Chaytor & Scmitter-Edgecombe, 2003). These are the verisimilitude and the veridicality approaches. The *verisimilitude* approach refers to the degree to which the cognitive demands of a test theoretically resemble the cognitive demands present in the naturalistic task environment. This approach has lead to the development of new assessment instruments aimed at capturing "the essence of everyday cognitive skills" (Chaytor & Scmitter-Edgecombe, 2003, p. 182) (e.g., Test of Everyday Attention) (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1996). *Veridicality* approach, on the other hand, refers to the degree to which the instrument's results can be related empirically to measures of everyday task performance. This approach relies on statistical techniques to study the relationship between performance on cognitive-perceptual tests and measures of daily life task functioning (e.g., ADL). Such correlative studies have demonstrated different results, but most indicate no more than low to moderate correlations (Bouwens et al., 2008; Chaytor & Scmitter-Edgecombe, 2003; Donkervoort, Dekker, & Deelman, 2002; Edmans & Lincoln, 1990; Korpelainen, Niilekselä, & Myllylä, 1997; Nygård, Amberla, Bernspång, Almikvist, & Winblad, 1998; Sveen, Bautz-Holter, Sødring, Wyller, & Laake (1999); Titus, Gall, Yerxa, Roberson, & Mack, 1991). The limitation of both of these approaches is that neither evaluates

neurobehavioral impairments directly in the natural context of daily life task performance.

Comparison of persons with RCVA and LCVA

Other limitations of existing research exploring the relation between ADL ability and neurobehavioral impairments among persons with RCVA and LCVA include (a) that comparisons are, at times, made by use of scales that include items from different constructs (Granger, Hamilton, & Fiedler, 1992; Yavuzer, Küçükdeveci, Arasil, & Elhan, 2001), as many existing scales include for example, body functions and ADL on same scale (Salter et al., 2005); (b) examination of limited number of a limited range of impairments such as only motor functions or only one ADL task such as locomotion (Goto et al., 2009); (c) exclusion of some persons from one of the groups due to specific impairments such as aphasia (Glymour et al., 2007); or (d) comparison being made based on scales that do not have known measurement properties, in contrast with the more resent trend within the field of rehabilitation where increased emphasis is placed on use of linear scales that have measurement properties (Haigh et al., 2001; McAllister, 2008; Tesio, 2003). This underlines that the comparisons are complicated and may lead to unintentional misinformation.

Development of the A-ONE placed in the context of Yerxa's model of an integrated profession

In addition to conceptual and process models that I mentioned at the beginning of this Introduction, one can consider models of the profession. One such model was described by Yerxa (1994), which she called a *model of an integrated profession*. This model depicts a circular flow from practice to ideas, then research, and finally, education and back again to practice. Thus, practice is considered to be both the source and the destination for the profession's ideas and research. A balance between all four components of an integrated profession model (practice, ideas, research, education) is necessary if we are to succeed and gain the needed flexibility for meeting the changing demands of society and the healthcare system.

Because I have used Yerxa's (1994) integrated profession model as a framework both in describing the original development of the A-ONE, but also as I have planned and implemented the research presented in this thesis, I review that model in relation to A-ONE development in the sections below. Specifically, I will review how the ideas for the A-ONE came from practice and resulted in developing the conceptual background of the A-ONE

and implementing research studies that led to its development as an ordinal ADL and NB scale. Then I will discuss briefly how the contributions of this development was published and integrated back into practice from a historical perspective. Because this cycle, shown in Figure 1, pertains to the history of the A-ONE, I call it an historical cycle.

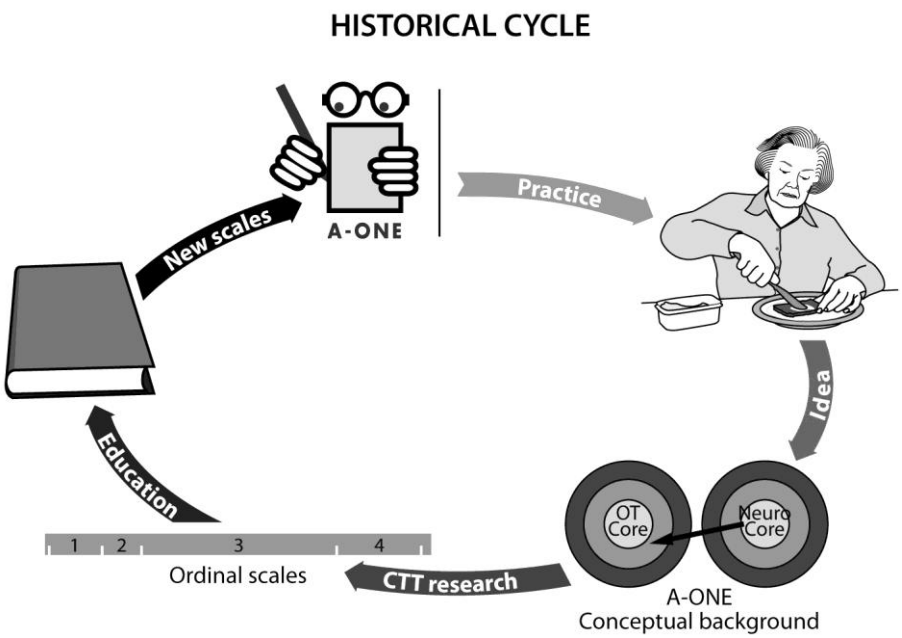

Figure 1. The circular path of the model of integrated occupational therapy profession: First, historical round for the A-ONE.

Practice as source of ideas

While practicing occupational therapy and performing ADL assessments, I noticed that we often obtain information not only about a person's independence level and assistance needed for ADL task performance, but also, through clinical reasoning, understanding of which underlying impairments are interfering with ADL task performance and restricting independence. For example, the person sitting beside me in the case example I presented earlier had been diagnosed with LCVA. When I

evaluated her, I observed that she could not put on a shirt. At the same time, I detected errors that I related to underlying ideational apraxia interfering with her ADL task performance. Thus, the clinical ideas that led to the development of the A-ONE were that I could gather information through observation of people's ADL task performances, and then combined with further reasoning about the function of the central nervous system (CNS), I could ultimately develop goals for intervention, select appropriate interventions, and provide education to the person, his or hers family or friends, and other members of the rehabilitation team.

Research, step 1: Developing a conceptual model for the A-ONE

The concepts used and defined, both conceptually and operationally in the A-ONE, were derived from within the discipline of occupational therapy, as well as from a comprehensive review of the literature related to neurology and neuropsychology (Árnadóttir, 1990). Two important global concepts were neurobehavioral impairment and occupational performance error. A *neurobehavioral impairment* was defined as a disorder of neuronal processing (e.g., disorders of body scheme, cognition, emotion, gnosis, language, memory, motor movement, perception, personality, praxis, sensory awareness, spatial relations, visuospatial skills), that ultimately was manifested as defective ADL task performance (Árnadóttir, 1990, 2004). More specifically, the concepts of neurobehavior and neurobehavioral impairment were linked to occupation, as elements of neurobehavior include different types of sensory stimuli evoked by different daily life tasks. These stimuli were viewed as being processed by different mechanisms within the CNS and resulting in different types of behavioral responses (e.g., affect, movement). Neurobehavior, therefore, was conceptualized as including the different types of pertinent neurological body functions necessary for performing ADL tasks, and when these functions become impaired, they are manifested as occupational errors (Árnadóttir, 1990, 2004b).

Occupational performance error, referred to in the remainder of this thesis as *error,* was defined as any deviations from flawless responses when performing ADL tasks (Árnadóttir, 2004b, 2009). Indications of neurobehavioral impairments that limit ADL task performance are based on detection of errors through task analysis of the observed ADL performance. *Task analysis* refers to a process where the cause of problems with an observed task performance is interpreted in relation to underlying body functions, person factors, or environmental factors (Fisher, 2009).

In the process of selecting which neurobehavioral items to be used in the A-ONE and developing operational definitions of those concepts, I used both activity and task analyses. *Activity analysis* is defined as the process of examining tasks in detail, by breaking them into their component parts, in order to understand and evaluate them (Árnadóttir, 1990; Llorens, 1986). Activity analysis can be based on particular theories or conceptual frameworks, and can be focused on specific body functions and/or the environment (Crepeau & Boyt Schell, 2008; Trombly Latham, 2008b).

During the development of the A-ONE, observable errors were subsequently classified and operationally defined as neurobehavioral impairments to be included in the A-ONE test manual (Árnadóttir, 1990, 2009). More specifically, the theoretical concepts gathered from the literature (e.g., ideational apraxia) subsequently evolved into relational statements that proposed a relationship between different factors — the ability to perform ADL tasks, neurobehavioral impairments, and the related functions of the CNS. That is, by forming statements explaining how neurobehaviors are related to daily activities, the second level of theory development, termed factor relating type of theory (Dickoff, James, & Wiedenbach, 1968) was reached. Specific examples of the factor-relating statements include: (a) performance of ADL tasks requires adequate functioning of specific parts of the central nervous system (CNS), and (b) impairment of certain body functions may result in dysfunction or inability to perform specific aspects of ADL tasks. More specifically, if we consider the woman in the case example presented earlier who had had a massive posterior parietal lobe lesion in the left hemisphere (CNS damage), and we consider that she tried to use a spoon to scoop milk from a glass (problem with ADL task performance), we can hypothesize, using clinical reasoning, that this type of error was related to ideational apraxia (neurobehavioral impairment). A small precentral lesion in the primary motor area (CNS damage) of another person may cause muscle paralysis of the contralateral side (neurobehavioral impairment), but not ideational apraxia. The paralysis, in turn, might be hypothesized as the reason the person was unable to use both hands to close fastenings on clothings or open containers (problems with ADL task performance). Thus, presence of neurological impairments can be hypothesized from observed errors during the person's engagement in ADL task performance.

In summary, the A-ONE was developed based on ideas obtained from practice, and that lead to the development of relational statements that were the conceptual basis for a new evaluation approach — one that would enable occupational therapists to not only evaluate the amount of assistance a

person needed with ADL task performances, but also, using the clinical reasoning described above, to determine what types of neurobehavioral impairments caused the person to need assistance. The conceptual model of the A-ONE cannot be claimed to be a conceptual practice model. Rather, the conceptual model was developed to be used as a guide to clinical reasoning when determining what underlying neurobehavioral impairments are the reason for a person's need for assistance during ADL task performance.

Research, step 2: Implementing validity and reliability studies of the A-ONE

The relational statements in the conceptual model of the A-ONE led to research questions, and studies were conducted to validate some of the relational statements. Research with the A-ONE has included a number of studies related to content validation, interrater reliability, inter-item correlations, factor analyses, concurrent and construct validation, and the results of these studies and the evidence they provide is summarized in Table 1.

Back to practice through education

The knowledge obtained from the studies summarized in Table 1 were subsequently returned back to the practice through publications (Árnadóttir, 1990; Nuwer, Árnadóttir, Martin, Ahn, & Carlson 1994; Steultjens, 1998; Gardarsdóttir & Kaplan, 2002) and training of therapists from different countries to use the A-ONE.

Since the A-ONE was developed, A-ONE courses have been offered in 11 different countries. These courses served as a mechanism to disseminate knowledge gained through research back into practice. In these training courses, occupational therapists learn how to observe and score the A-ONE reliably. Feedback from occupation therapists who have taken these courses have often focused on the value of the A-ONE to practice but also the fact that it was limited as it could not be used to generate an overall measure of a person's ADL ability nor a measure of the extent to which neurobehavioral impairments interfere with ADL task performance. Thus, while the scales were useful to occupational therapists, they were limited in terms of sharing the results of the A-ONE to others or evaluating change following intervention. It was this feedback from practice that has resulted in my reentering the circle described in Yerxa's (1994) model in the hopes of being able to apply Rasch measurement methods and realize the aims of this thesis.

Aims of this Thesis

The overall aim of this doctoral thesis was to further develop and validate the A-ONE, with a primary focus on exploring the potential for converting the ordinal ADL and NB scales to interval scales. This included examination of which of the neurobehavioral items would be most beneficial for constructing a unidimensional, linear Neurobehavioral Impact (NBI[2]) scale for evaluating persons with different types of neurological diagnosis and further validation of the NBI scale by examining for differences in NBI measures between persons with RCVA and LCVA. The specific aims are listed below.

Specific aims

- To use Rasch analysis methods to explore the rating scale structure, and aspects of scale validity and reliability of the ADL scale of the A-ONE (Study I).

- To examine underlying nonmotor neurobehavioral impairments to verify if such impairments can be viewed as unidimensional when evaluated in the context of performance of ADL (Study II).

- To determine if a single NBI scale could be constructed for use across different diagnostic groups, and, if not, can deficit-specific scales be developed (Studies II and III).

- To validate the NBI scale by exploring if persons with RCVA and LCVA differ significantly in NBI measures when ADL ability is controlled (as a covariate) (Study IV).

[2] Neurobehavioral Impact (NBI) scale is the name given to the Rasch analyzed version of the A-ONE Neurobehavioral scale, referred to in earlier sections of this thesis as NB scale.

Methods

Participants

The data used in all four studies were gathered retrospectively from A-ONE evaluation forms in hospital records from two different wards (rehabilitation and geriatric) at Landspítali University Hospital in Iceland. Except for the first study of the ADL scale, where data for three persons with other neurological diagnoses were included, only records from persons diagnosed with CVA and dementia were analyzed. Study IV included only persons with CVA.

For Studies I and II, data from 209 and 206 records, respectively, for persons who had been hospitalized between 2000 and 2004, were analyzed. Our goal was to have sample sizes of at least 30 in each diagnostic group as this number would be sufficient to provide a 95% confidence interval for the stability of the estimated items difficulty calibrations within an absolute value of 1.0 logit (Linacre, 1994).

For Studies III and IV, the data was expanded by adding records from persons hospitalized between 1994 and 1999 and 2005, so that the sample now included persons hospitalized between 1994 and 2005. This resulted in data from 422 records included in the analysis for Study III, 200 from persons diagnosed with dementia and 222 from persons diagnosed with CVA. Study IV included data from 115 records of persons diagnosed with CVA, these being selected from the 222 records available for Study III.

Different diagnostic groups were included in each of the four studies, but heterogeneity of both ability and diagnoses, reflecting the group of persons to which the A-ONE is intended to be applied, was kept in mind in the selection of participants. While there were limited data available for persons in the acute phase post CVA or for persons in later stages of dementia, there was a spread of ability within all four groups across the range of the different A-ONE scales. See Table 2 for demographic characteristics and further details of the sample diagnostic composition of participants included in Studies I–IV.

Table 2 Demographic Characteristics and Sample Diagnostic Composition of Participants in Studies I–IV

Study	***n***	**Diagnoses** (*n*)	**Gender** *n* (% men)	**Age** (years) *M*±*SD*, (range)
I	209	**CVA** - RCVA (37*) - LCVA (36*) - Other CVA (22*) **Dementia** - DAT (40*) - DU (71*) - Other (3*)	84 (40.2)	73.7±13.0 (22–99)
II	206	**CVA** - RCVA (37*) - LCVA (36*) - Other CVA (22*) **Dementia** - DAT (40*) - DU (71*)	83 (40.3)	74.9±12.7 (22–99)
III	422	**CVA** - RCVA (37*) - RCVA (71**) - LCVA (36*) - LCVA (78**) **Dementia** - DAT (40*) - DAT (31**) - DU (71*) - DU (58**)	187 (44.3)	73.0±13.7 (22–99)
IV	215	**CVA** - RCVA (37*) - RCVA (66**) - LCVA (36*) - LCVA (76**)	129 (69)	66.53±14.0 (22–91)

Note. RCVA = Right cerebrovascular accident, LCVA = Left cerebrovascular accident, DU = Dementia unspecified, DAT = Dementia Alzheimer type, Other CVA = other types of CVA, Other = other diagnoses.

* Data collected from charts of persons hospitalized between 2000–2004.

** Additional data collected from charts of persons hospitalized between 1994–1999 and 2005.

Instrumentation

As summarized in the Rationale, the A-ONE is composed of an ADL scale and a NB scale that, in turn, is comprised of two subscales, the NBSIS and the NBPIS. Each of the 22 items in the five different domains on the ADL scale of the A-ONE is rated using a 5-category ordinal rating scale: zero (full assistance needed), 1 (minimum assistance needed), 2 (verbal assistance needed), 3 (supervision needed), and 4 (independent). Persons are scored based on the observed level of assistance needed for the ADL performance, and are not penalized for using assistive devices. Table 3 includes items from the ADL scale.

Table 3 Domains and Original Items on the ADL Scale of the A-ONE

Dressing	**Transfers and mobility**
- Put on shirt	- Sit up in bed
- Put on pants	- Transfer from sitting
- Put on socks	- Maneuver around
- Put on shoes	- Transfer to toilet
- Manipulate fastenings	- Transfer to tub
Grooming and hygiene	**Feeding**
- Wash face	- Drink from glass/cup
- Comb hair	- Use finger to bring food to mouth
- Brush teeth	- Bring food to mouth by fork or spoon
- Shave beard/apply cosmetics	- Use knife to cut and spread
- Perform toilet hygiene	
- Bathe	
Communication	
- Comprehension	
- Expression	

The NB scale is used to evaluate the extent to which consequences of neurological dysfunction (neurobehavioral impairments) impact ADL task performance, and the NB items are scored by use of operational definitions of neurological impairments included in the test manual (Árnadóttir, 1990, 2004a, 2009). More specifically, the occupational therapist uses the NB scale to record the detected presence or absence of neurobehavioral impairments impacting ADL. Most of the 46 NBSIS items are independently rated more

than once in connection with performance of different ADL tasks (e.g., unilateral body neglect–dressing, unilateral body neglect–grooming and hygiene, unilateral body neglect–transfers and mobility, and unilateral body neglect–feeding). When errors that reflect the presence of neurobehavioral impairments are observed, the therapist specifies whether they interfere with ADL task performance, and subsequently, the type of assistance needed to overcome the errors during the performance. That is, the 5-category rating scale of the NBSIS reflects the type of assistance needed to overcome the performance errors during ADL task performance (0 = no problem observed, 4 = maximum physical assistance required). In contrast, the 31 NBPIS items are only rated as present or absent, based on an observed error during performance of at least one ADL task (Árnadóttir, 1990).

Some of the items of the NBPIS require additional administrative procedures, and thus, are not scored based on direct observation of ADL task performance. Some of the other neurobehaviors are scored more than once on redundant items. Finally, there are other items that are rarely used with this population. Thus, the following inclusion criteria for items was set for this study: (a) only items scored based on observation of errors in natural context of ADL task performance, (b) absence of redundancy between items, (c) observed error ≥ 10. Table 4 includes both items included and excluded in studies of the NB scale of the A-ONE.

When the different ADL domains and neurobehavioral impairments that might affect ADL performance included in the A-ONE are scored, the occupational therapist (a) fills in scores for the level of assistance needed for ADL task performance (i.e., scores items on the ADL scale); (b) writes comments in the Comments and Reasoning sections of the score form about the ineffective actions observed as errors during ADL task performance; (c) reasons, based on the content of the observed errors, about the type of impairment responsible for the error; (d) scores the respective neurobehavioral impairment item based on the type of assistance level needed to complete the ADL task (NBSIS), or whether the impairment is present or not (NBPIS).

It is important to again stress that the A-ONE NB scale was not developed to evaluate presence of neurobehavioral impairments per se. Rather, the occupational therapist uses it only to identify which neurobehavioral impairments are impacting ADL task performance when ADL task performance is observed in naturalistic contexts. See Figure 2 for a scoring sample from the Dressing domain of the ADL scale and pertinent NBSIS items for the person with eating and dressing problems described in the case sample earlier in the Rationale and Introduction of this thesis. Finally, a

Table 4 Items on the Original Neurobehavioral Scale Included and Excluded before Rasch Analyses

Items included in analyses	Items excluded before analysis
Specific impairments	
- Motor apraxia†	- Somatoagnosia**
- Ideational apraxia†	- Jargon aphasia*
- Unilateral body neglect†	
- Spatial relations†	
- Unilateral spatial neglect†	
- Organization and sequencing†	
- Motoric—right side†	
- Motoric—left side†	
- Perseveration‡	
- Topgraphical disorientation*	
- Sensory aphasia*	
- Anomia*	
- Paraphasia*	
- Expressive aphasia*	
- Dysarthria*	
Pervasive impairments	
- Lability*	- Astereognosis*
- Apathy*	- Motor impersistence*
- Depression*	- Visual object agnosia*
- Irritability*	- Visual spatial agnosia*
- Frustration*	- Associative visual agnosia*
- Restlessness*	- Anosognosia*
- Insight*	- Body part identification*
- Judgment*	- Right/left discrimination*
- Confusion*	- Concrete thinking*
- Attention*	- Euphoria*
- Distraction*	- Agressiveness*
- Initiative*	- Alertness*
- Motivation*	- Absentmindedness*
- Performance latency*	- Disorientation*
- Working memory*	- Long term memory*
- Confabulation*	

† NB items scored four times, once each in Dressing, Grooming and hygiene, Transfers and Mobility, and Feeding domains

‡ NB item scored five times, once each in Dressing, Grooming and hygiene, Transfers and mobility, Feeding, and Communication domains

* NB item scored only once

** NB item scored twice, once each in Dressing and Grooming and hygiene

PRIMARY ADL ACTIVITY	SCORING					COMMENTS AND REASONING
DRESSING	**IP SCORE**					
Put on shirt	4	3	2	(1)	0	*Planning action, coordinating hands: Cannot figure out how to start or continue. Paralyzed and needs one handed-technique.*
Put on pants	4	3	2	1	(0)	*Planning, stabilizing body, bending down: Cannot figure out how to start or continue. Needs one handed-technique. Unstable when standing.*
Put on socks	4	3	2	1	(0)	*Planning, manipulating (one handed-technique), bending: Finger movements left hand.*
Put on shoes	4	3	2	(1)	0	*Planning, reaching, bending, manipulating: Able to get left foot into slip-on shoe. Full assistance with right shoe.*
Manipulate fastenings	4	3	2	(1)	0	*Manipulating, coordinating buttons.*

NB IMPAIRMENT	NB SCORE					COMMENTS AND REASONING
Motor apraxia	0	1	2	(3)	4	*Manipulate, grasp/turn socks, shirt and pants correctly.*
Ideational apraxia	0	1	2	3	(4)	*Planning location of arms in shirt, and actions for socks and pants using one-handed techniques.*
Unilateral body neglect	(0)	1	2	3	4	
Somatoagnosia	(0)	1	2	3	4	
Spatial relations	(0)	1	2	3	4	
Unilateral spatial neglect	(0)	1	2	3	4	
Abnormal motor function: Right	0	1	2	(3)	4	*Limited ability to use one-handed techniques.*
Abnormal motor function: Left	(0)	1	2	3	4	
Perseveration	0	1	2	(3)	4	*Several buttons matched with one buttonhole.*
Organization and sequencing	0	1	2	3	(4)	*Plan for putting on all garments, is missing. Sequencing of action steps is out of order.*

Figure 2. Scoring sample from the Dressing domain of the ADL scale and pertinent NBSIS items for the case sample with LCVA.

summary of the existing evidence for validity and reliability of the A-ONE was presented in the Introduction.

Procedures

All participants had been evaluated by using the A-ONE as a part of routine occupational therapy services at the Landspítali University Hospital. The number of occupational therapists who performed the evaluations ranged from 8–11 in the different studies (see Table 5). All the therapists had gone through a 5-day A-ONE training course and administered the evaluations according to the standardized procedures described in the A-ONE manual.

As this was a retrospective study, it was not possible to obtain information on rater reliability of the therapists involved. Prior research has supported acceptable levels of rater reliability for the ADL and the NB scales (Árnadóttir, 1990, 2004a). I extracted raw data from the participants' records.

Data analysis

In the following sections, I will present the methods used for analyzing the data. More specifically, I will start by presenting the methods we used for examining aspects of scale validity and reliability of the ADL scale, undertaken in Study I. Then I will proceed to presenting the methods we used for examining unidimensionality of various versions of the new NBI scale, explored in Studies II and III. Finally, I will present the methods we used for examining for significant differences in NBI measures between persons with RCVA and LCVA. A summary of the analyses performed is shown in Table 5.

The raw scores were analyzed using different versions of the WINSTEPS Rasch computer software program (Linacre 1991–2008) across the four studies. Additionally, the Statistical Package for the Social Sciences (SPSS) 12.0 was used for calculating demographic information and implementing other statistical analyses.

ADL scale

In the first phase of the data analysis for Study I, a Rasch rating scale analysis of the ADL item raw scores was implemented. We then proceeded to evaluate the psychometric properties of the 5-category rating scale using Linacre's (2002) guidelines and consideration of threshold disordering (Andrich, 1996). We were prepared to collapse non-advancing categories

with the ones below if the average measures did not advance with category (Linacre, 2002), provided the joining of adjacent categories made sense theoretically. Our goal was to ensure the best person separation along the variable when determining optimal categorization (Lopez, 1996).

Table 5 Overview of Data Collection and Analyses for Studies I–IV

Study	Data collection	Data analyses
I	*ADL scale* Eleven occupational therapists at rehabilitation and geriatric wards at LSH	- *Rasch measurement*: Rating scale model, examination of rating scale properties, item goodness of fit, PCA, item calibration values, targeting, separation index, reliability index - *Other statistics*: Descriptive statistical analyses
II	*NBI scale* Eleven occupational therapists at rehabilitation and geriatric wards at LSH	- *Rasch measurement:* Simple Rasch model for dichotomous data, examination of item goodness of fit, PCA, item calibration values, targeting, separation index, reliability index - *Other statistics*: Descriptive statistical analyses
III	*NBI scale* Eleven occupational therapists at rehabilitation and geriatric wards at LSH	- *Rasch measurement statistics*: Simple Rasch model for dichotomous data, examination of item goodness of fit, PCA, DIF, DTF, item calibration values, targeting, separation index, reliability index - *Other statistics*: Pearson product moment correlations, descriptive statistical analyses
IV	*NBI scale* Eight occupational therapists at rehabilitation wards at LSH	- *Rasch measurement*: Simple Rasch model for dichotomous data (NBI-CVA scale) - *Other statistics*: Pearson product moment correlations, ANCOVA, descriptive statistical analyses

Subsequently, we examined for internal scale validity by examining the *MnSq* and *z* goodness-of-fit statistics for the ADL items. We evaluated both infit and outfit statistics. Infit statistics are weighted to give more value to on-target observations. Thus, infit statistics are sensitive to item performance and are more informative when exploring internal scale validity (Bond & Fox, 2007; Wright & Masters, 1982). Outfit statistics are unweighted estimates of the degree of fit of the responses. Outfit statistics tend to be influenced by off-target observations such as outlying person responses (Bond & Fox, 2007). Because our focus was primarily on

examining internal scale validity of the ADL scale, we chose to focus on infit (Bond & Fox, 2007; Wright & Masters, 1982) revealing information on item performance.

The expected *MnSq* value is 1.0 (Bond & Fox, 2007). High *MnSq* values signal unexpectedly high or low scores, whereas low *MnSq* values indicate overly predictable score strings and the failure of the item to provide independent information about the status of the person (Bond & Fox, 2007; Linacre et al., 1994; Wilson, 2005; Wright, 1995). We chose to focus on high *MnSq* values, as these are a particular threat to validity (Wilson, 2005; Wright, 1995). Our criteria for failure to meet the assertions were based on the combined consideration of *MnSq* > 1.4 (Wright & Linacre, 1994) and standardized $z \geq 2$ (Wilson, 2005). However, we considered the theoretical importance of items before item removal (Bohling, Fisher, Masters, & Bond, 1998), acknowledging that up to 5% of items on a scale are expected to misfit by chance (Smith, 1991). We chose this more conservative approach as one, which would enable us to consider an item's future potential to be revised (e.g., split into two or more new items), rather than immediate item removal before further research could be implemented (Linacre, 1995).

We also evaluated unidimensionality via PCA. If the proportion of variance explained by the measures (Rasch dimension) was > 60%, and the proportion of unexplained variance accounted for by the first contrast (the largest secondary dimension) was < 5%, the results would be considered to support unidimensionality (Linacre, 1991–2006). Additionally we verified the logical ordering of items along the scale and targeting of the items to the abilities of the persons. Exploration of targeting included comparison of means and ranges of the item and person distributions, as well as exploration of any gaps along the hierarchical continuum.

Neurobehavioral scale

As the NBSIS has a 5-category rating scale and the NBPIS has dichotomous items, we determined it to be clinically most practical to dichotomize all items for analysis in Studies II and III. Thus, the NBSIS rating scale items were recoded in the same way as the NBPIS items are scored (i.e., present or absent), and the simple Rasch model (a model for dichotomous data) was used for the analysis. The same criteria for acceptable goodness of fit statistics and PCA were used in Studies II and III as for Study I, except that both infit and outfit statistics were considered equally important for item omission and no misfit was allowed. Thus, the item omission criteria were more stringent for examination of the NB items than we used earlier for the ADL scale.

The NB scale analyses progressed in two phases. In one phase, the potential for forming a common unidimensional global neurobehavioral impact (NBI) scale (referred to here as the global NBI scale) using (a) all NB scale items, except the non-motor neurobehavioral items, that met the inclusion criteria (Study II); and (b) all items that met the inclusion criteria, including the motor items, was explored. When the motor items were included, motor items for the right and the left body side were collapsed into one motor item for each domain (Study III). In the other phase of the analyses, the potential for forming unidimensional hierarchies for diagnoses-specific groups (referred to here as diagnosis-specific NBI scales) from (a) all NB scale items, except the neurobehavioral motor items, that met the inclusion criteria (Study II); and (b) all items that met the inclusion criteria, including the motor items, was explored (Study III).

In Study II, we first attempted to form a common global hierarchy, and then progressed to evaluating for three diagnosis-specific NBI scales based on splitting the sample into three different diagnostic subgroups (RCVA, LCVA, dementia). For these analyses, we selected only those participants with LCVA, RCVA, or dementia to create three subgroups, LCVA (n = 36), RCVA (n = 37), and dementia (n = 111). Since the sample size used for exploring the potential for diagnoses-specific NBI scales in Study II was too small to reveal reliable results (Linacre, 1994), Study III was undertaken with almost double the sample size. In Study III, the first phase involved evaluating for four diagnostic-specific NBI scales, one each for RCVA, LCVA, dementia of Alzheimer type (DAT), and unspecified dementia (DU); and in the second phase of Study III, we concentrated on a search for an item subset that could be used to construct a single unidimensional global NBI scale that could be used across all four diagnostic groups.

In Study III, the evaluation of differential item functioning (DIF) also was added to the examination of unidimensionality. Then, to evaluate for differential test functioning (DTF), person measures for each of the different diagnoses-specific NBI scales, as well as for the common short form NBI scale, were obtained through Rasch analysis and compared. That is, we evaluated if, despite statistical DIF, the common NBI scale could be determined to be fair to persons with different diagnoses.

In addition to the NBI scales reported in Study III, in Study IV we also used similar procedures to develop a 53 item common global NBI scale that could be used with persons with either RCVA or LCVA (NBI-CVA scale). The same methods and criteria for acceptable psychometric properties were applied as were used in Study III.

Difference in NBI measures between right and left CVA

The initial sample for Study IV was comprised of all persons with RCVA or LCVA that had been included in Study III (n = 222). In the first phase of the study, we used the ADL scale developed in Study I and the NBI-CVA scale to generate ADL and NBI measures for all participants with RCVA and LCVA. Seven persons with maximum scores were omitted after the ADL measures were generated, and, therefore, NBI measures were obtained only for the remaining 215 persons (103 RCVA and 112 LCVA). In the second phase of Study IV, we calculated the Pearson product moment correlation (r) between the ADL and NBI measures for the entire group was computed to verify our assumption that the two scales were evaluating different but overlapping constructs. The following criteria were used to classify the strength of the relation: r = zero–0.30 = little if any; 0.30–0.50 = low; 0.50–0.70 = moderate; 0.70–0.90 = high; 0.90–1.00 = very high correlation; we expected a moderate relation.

Subsequently, in the third phase, analysis of covariance (ANCOVA) was used to examine if the mean NBI measures of the RCVA and LCVA groups differed significantly ($p \leq .05$). ADL ability was used as the covariate to control for potential differences between groups in ADL ability.

Ethical considerations

Prior to collection of raw scores and participant demographic information from the available A-ONE forms in the hospital records, written approval for the study was obtained from the Ethical Committee of Landspítali University Hospital (Study number 36/2003). As this was a retrospective study, use of the data did not affect the participants in any way, and all personal identification was removed from the data.

Results

ADL scale

Psychometric properties of the rating scale

Results from the Winsteps analysis of the psychometric properties of the rating scale revealed that the criteria for category frequencies, outfit *MnSq,* and category measures were met. However, some threshold disordering was detected. The disordering was eliminated in subsequent analyses by successful combination of two categories, verbal assistance (score = 2) and supervision (score = 3).

Internal validation of the ADL scale

Internal scale validation was performed by several methods. These included examination of goodness of fit for items, logical hierarchical ordering of items, and targeting, as well as PCA analysis.

Goodness of fit and PCA

When analyzing the 22 items from the five original domains of the A-ONE, the two items from the communication domain ("Expression" and "Comprehension") misfit the model. The item of "Use of knife" from the feeding domain also demonstrated infit misfit. As this resulted in a total of 13.6% item misfit to the model, beyond the predetermined criteria of accepting 5% misfit, another analysis was run after the two communication items were removed. This resulted in continued misfit of the item "Use of knife" ($MnSq > 1.4$, $z \geq 2$), but all other items continued to demonstrate acceptable infit goodness of fit. As the item "Use of knife" was determined to be of important content for the construct, and total item misfit was within the predetermined criteria, the item was retained. PCA of the 20 remaining items confirmed unidimensionality of the items forming the ADL scale.

Examination of hierarchical item order and targeting

The hierarchical order of item difficulty appeared to be logical, with three of the four feeding items being the easiest items, and items from the Transfers and mobility domain ("Transfers to tub") and Grooming and hygiene domain ("Bathe") being the hardest items. When exploring targeting of person ability to item difficulty, there was a discrepancy of 1.61 logits between the mean measures, indicating that the items might not be well

targeted to the most able persons. Nine persons presented with maximum scores. The only gap detected to exceed 0.50 logits was at the lowest performance categories, but no persons in the sample were located across from this gap.

Reliability

The person separation index increased slightly, by both the collapsing of rating scale categories and removal of the communication items. The final analysis revealed a person separation index of 2.93 and separation reliability coefficient of 0.90. These results indicate that we can reliably differentiate the sample into at least three statistically distinct strata of ADL ability. The item separation index was 8.02 with an associated reliability coefficient of 0.98.

Neurobehavioral Impact (NBI) scales

Global Neurobehavioral Impact scales

Three global NBI scales were explored. The first included only non-motor items (Study II). Four of the original 50 items were omitted (anomia, expressive aphasia, working memory, and motivation) as they failed to demonstrate acceptable goodness of fit. The remaining 46 items failed to reach the predetermined criteria for the PCA; only 56.8% of the variance was explained by the Rasch factor.

The analysis for the second global NBI scale (Study III) included the collapsed motor impairment items, and the results revealed that 33 items were "common" to all four diagnostic groups (see below), and, therefore, had the potential to be retained in the second global NBI scale. Simple Rasch analysis of those common items resulted in further omission of four items due to misfit. The 29 remaining items demonstrated acceptable goodness of fit to the assertions of the simple Rasch model (see Table 6 for summary of the psychometric qualities of the global common short-form NBI scale developed in Study III. Also included in Table 6 are the results of the analysis of a third global combined NBI-CVA scale developed and used in Study IV. The results for the first global NBI scale developed in Study II are not included in Table 6.

Subsequent PCA analysis of the 29 items remaining on the second global NBI scale revealed a Rasch factor explaining 72.8% of the variance which supported unidimensionality. A number of persons had extreme maximum or minimum scores. No items had extreme measures. The mean person measure was -1.74 (*SD* = 1.34), indicating that there were

considerably more difficult items to evaluate persons of low ability, but very few items to evaluate persons of higher ability. The item-person map supported this finding, and indicated a lack of items for evaluating persons with only higher level, milder neurobehavioral impairments that impacted ADL task performance, and only a few retained items targeted to the persons. There were also large gaps in the hierarchy (see Figure 3). The gaps underline that items of the difficulty level needed to evaluate persons located at these ability levels are missing from the hierarchy. The person separation index indicated that the sample can be separated, with reasonable reliability, into at least two statistically distinct strata of neurobehavioral ability. Most of the items displayed DIF for almost 50% of the group comparisons.

The NBI-CVA scale used to generate person measures in Study IV displayed acceptable goodness of fit statistics for all retained items (both infit and outfit), $MnSq \leq 1.4$ and $z < 2$, and acceptable PCA. As shown in Figure 4 the NBI-CVA scale demonstrated better overall targeting than did the common short-form global NBI scale.

Table 6 Summary of Findings based on Rasch Analyses of Different NBI Scales

	Diagnosis-specific scales				Global scales	
	RCVA scale	LCVA scale	DAT scale	DU scale	CVA scale	Common-scale
Persons (*n*)	108	114	71	129	215	422
Items (*n*)	51	42	49	40	53	29
PCA: Rasch factor	89.7%	91.1%	88.8%	98.1%	79.2%	72.8%
PCA: First contrast	1.6%	1.3%	1.0%	0.2%	1.7%	4.5%
Person separation index	2.57	1.94	1.93	2.54	2.20	1.64
R (persons)	0.87	0.79	0.79	0.89	0.83	0.73
Mean person measure	-1.67	-2.07	-2.48	-2.27	-1.44	-1.74
Mean person *SE*	0.51	0.63	0.64	0.66	0.46	0.70
N of persons with maximum or minimum scores	0	9	0	3	6	30

Note. CVA = Cerebrovascular accident, RCVA = Right CVA, LCVA = Left CVA, DAT = Dementia Alzheimer type, DU = Dementia unspecified, Common-scale = RCVA, LCVA, DAT, DU.

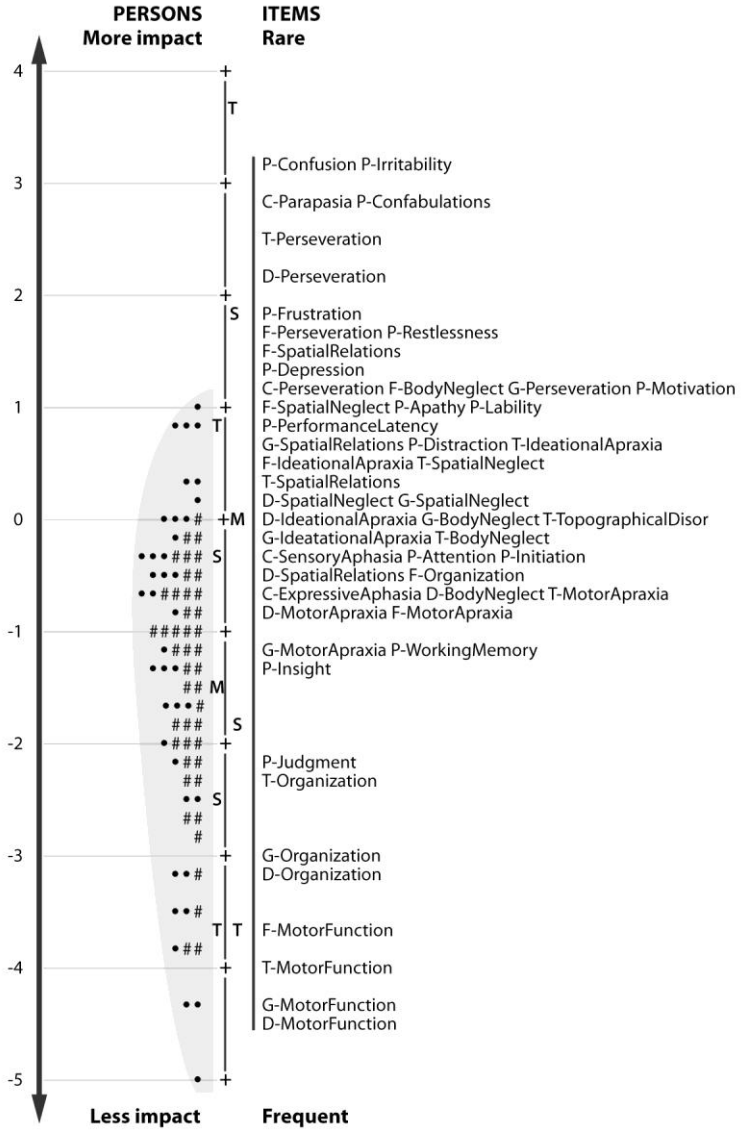

Figure 3. Common short-form NBI scale: Item-person map indicating spread of items and people along the logistic scale.

=4 persons

• =1 person

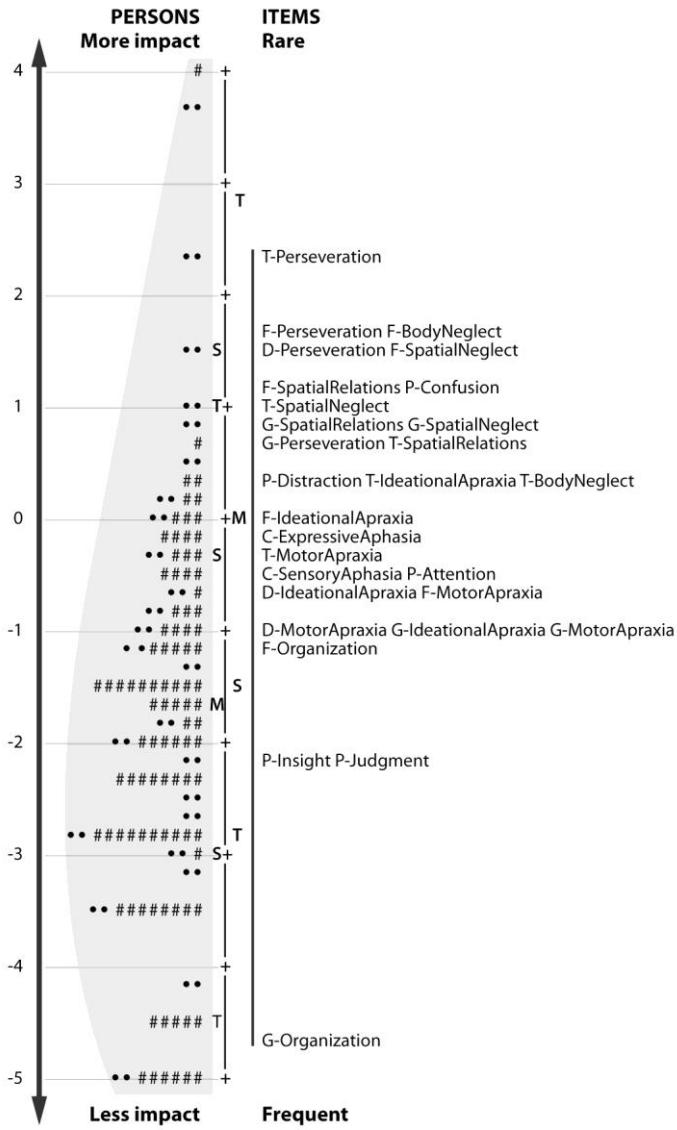

Figure 4. NBI-CVA scale: Item-person map indicating spread of items and people along the logistic scale.

= 4 persons

• = 1 person

Diagnosis-specific Neurobehavioral Impact scales

In Study II, we initiated exploration of the possibility for creating diagnostic-specific NBI scales using the data from 206 persons, but without inclusion of motor NB items. After removing misfitting items one at a time for each of the diagnostic groups, we were able to construct three unidimensional diagnosis-specific NBI scales. The items included in the final NBI scales varied among diagnostic groups.

When we expanded our analysis in Study III to include data for 422 persons classified into four different diagnostic subgroups and 55 items, we were able to construct four psychometrically sound diagnosis-specific NBI scales (see Table 6). The number of omitted items due to unacceptable infit and/or outfit statistics ranged from four to 15. Table 7 summarizes the number of items retained and omitted in the different analyses of all NBI scales constructed and/or used in Studies II−IV. The results of the PCA analyses for all diagnosis-specific NBI scales confirmed unidimensionality. As the diagnoses-specific NBI scales developed in the second study were viewed as preliminary in relation to those developed in Study III, the specific details of the results of the Rasch analyses are not included in Table 6.

Differential item and differential test functioning

When we evaluated for the presence of DIF in Study III, we found that most of the items displayed DIF for almost 50% of group comparisons. Since analysis of DIF in Study III confirmed our expectation that DIF would be present, we proceeded to implement comparisons of the paired person NBI measures (diagnosis-specific vs. common short-form scale) for each group to evaluate for DTF. The correlations coefficients ranged from r = .90 to r = .99). All participants except one person with RCVA (1/108 persons with RCVA, 1/422 persons in entire sample) had NBI measures within the 95% confidence interval control lines, supporting that presence of DIF did not result in DTF.

Difference in NBI measures between persons with RCVA and LCVA

Since we found an expected moderate correlation (r = -0.57) between person measures from the ADL scale and the NBI-CVA scale, we proceeded, as planned, to use ANCOVA procedures to determine if there is a difference in the extent to which neurobehavioral impairments impact ADL task performance between persons with RCVA and LCVA. The results revealed that the two groups do not differ in extent of impact of neurobehavioral

impairments impacting ADL as reflected by no significant difference in NBI-CVA measures between groups (F [1, 212] = 2.910, p=0.090).

Table 7 Overview of Analyzed and Omitted Items on Scales Created in Studies I–IV

Study	Scale	Total items	Items analyzed	New Rasch scale	Misfitting items	Items in final scale
I	FI	22	22	ADL	3	20
II	NBI	77	50 *	Single global scale	4	46
				RCVA	3	47
				LCVA	2	48
				Dementia	3	47
III	NBI	77	55 **	Common short-form scale	26	29
				RCVA	4	51
				LCVA	13	42
				DAT	6	49
				DU	15	40
IV	NBI			Combined CVA	2	53

CVA = Cerebrovascular accident, RCVA = Right CVA, LCVA = Left CVA, DAT = Dementia Alzheimer's type, DU = Dementia unspecified, Common-scale = RCVA, LCVA, DAT, DU

* Inclusion criteria: Observation of ADL errors in natural context only, absence of redundancy, frequency of ≥10, no motor items; dichotomized data

** Same as above but including motor items collapsed for right and left body sides

Discussion

This thesis focused on the examination of whether the scales of the A-ONE could be converted to unidimensional interval scales used to generate valid measures of ADL ability and neurobehavioral impact (NBI scale) that had the potential to be used to measure change and compare groups, despite different types of neurological disorders or underlying neurobehavioral impairments. Thus, just as I, in the Introduction, described the original development of the A-ONE by following the circular path depicted in Yerxa's (1994) integrated profession model (see Figure 1), the studies included in this thesis can be viewed within the context of a second round along the circular path depicted in Yerxa's model (see Figure 5). In this second round, which I call a *measurement cycle*, data have been tested against the assertions of different Rasch models in order to begin to gather evidence that the newly constructed A-ONE scales are valid and reliable. I will, therefore, now discuss the value of the information we have obtained. First, I will discuss what new validity and reliability evidence has been generated, and how that evidence adds to that which already existed for the A-ONE. Second, I will discuss how this evidence might have potential to be returned to practice and applied in a clinically useful form.

New validity evidence: From CTT to MTT

As I discuss our findings related to validity evidence from the four MTT studies that comprise my thesis, I will structure my discussion using the five types of validity evidence included in the "Standards" (AERA et al., 1999) (see Table 8). Recently, several authors (Lim et al., 2009; Smith, 2001; Wolfe & Smith, 2007b) have attempted to relate the different types of validity evidence to the types of evidence that can be generated using Rasch analyses. Because they used different classification systems, their summaries are often confusing, and sometimes contradictory and/or redundant. I have, therefore, created a structure that helped me avoid confusion and overlap, yet remained consistent with the original AERA et al. classification. That structure is summarized in Table 8.

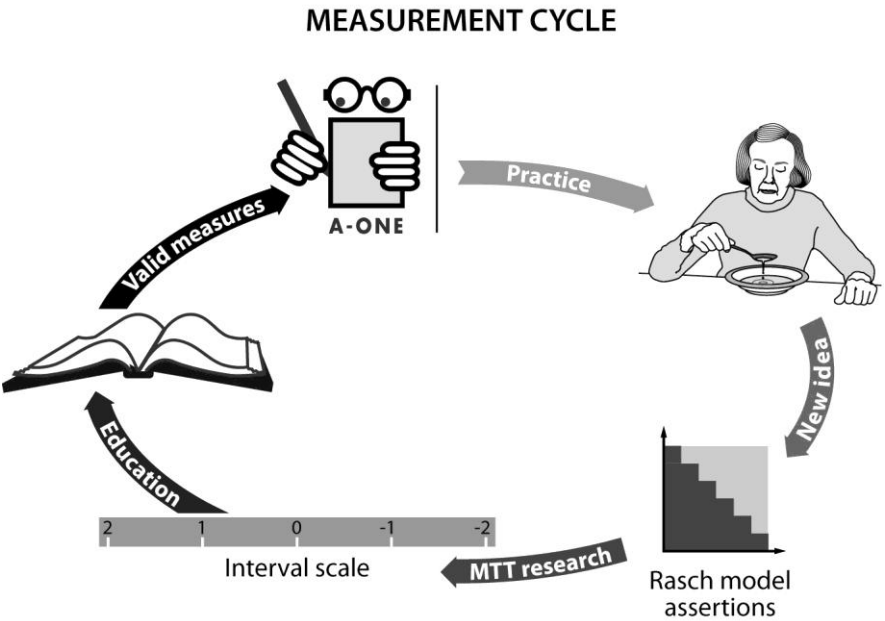

Figure 5. The circular path of the model of an integrated occupational therapy profession: A-ONE measurement cycle.

Evidence based on the content

In the older classification of validity (i.e., content, construct validity, and criterion), what is now called *evidence based on content* was referred to as content validity (Goodwin, 2002; Lim et al., 2009; Smith, 2001; Wolfe & Smith, 2007b). As noted in Table 1, I based the items included in the A-ONE on review of the literature and a previous study involving an expert panel provided evidence for content validity (Árnadóttir, 1990). It is clear, however, that the results of the Rasch studies described in this thesis have added to that evidence through information about the empirical item hierarchies and targeting of the items to the people tested.

Table 8 Different Types of Validity Evidence in Relation to Rasch Measurement Methods

Type of evidence	Methods
Evidence based on test content — Does the content of the test represent the domain it is proposed to measures in term of relevance, representativeness, and technical quality?	- Evaluation by expert panels - Evaluation of hierarchical ordering of items (item difficulty) - Evaluation of spread and range of items (targeting) - Evaluation for potential gaps in the item hierarchy (targeting)
Evidence based on response processes — Do the responses of the people tested fit the defined construct?	- Evaluation of the fit of the persons (person response validity; goodness-of-fit statistics for people) - Evaluation of the hierarchical ordering of the people (person measures)
Evidence based on internal structure — Do the internal components of the test combine to produce the desired effects?	- Evaluation of the psychometric qualities of the rating scale - Evaluation of item goodness of fit (internal scale validity) - Evaluation of unidimensionality using PCA - Evaluation for DIF
Evidence based on relations to other variables — Do the person measure relate to other variables in the manner expected?	- Evaluation of relationships to other test scores/measures - Evaluation of whether groups known to/not to differ, do/do not differ based on the test results
Evidence based on consequences of testing — Are there positive or negative effects related to testing (e.g., test bias, potential benefits)?	- Evaluation for test bias (i.e., DTF) - Evaluation of whether or not anticipated benefits were realized

ADL scale

The examination of the hierarchy of the ADL items (Study I) did provide content evidence in that the items theoretically and empirically expected to be easy or hard were found to be easy or hard, for example, when compared to the hierarchies of other existing ADL assessments. That is, while the construction of the A-ONE was not based on an expected item hierarchy, the obtained item hierarchy for the ADL scale can now be compared to hierarchies obtained from other similar instruments as well as to the developmental sequence of ADL ability. More specifically, by examining the Rasch analyzed item hierarchy for the ADL scale, it became evident that the hierarchy is in agreement with what would be expected based on the developmental sequence of ADL and Rasch studies of the FIM and the Barthel Index (Linacre et al., 1994; Tennant et al., 1996). Thus, the feeding items are the easiest items (except for "Use Knife" which is not a separate item on the FIM or Barthel) and items such as "Transfers to tub" and "Bathe" are the hardest ones.

Targeting was found to be less than ideal as the mean person ability measure (0.61 logit) was higher than the mean item difficulty calibration (zero logits). This means that the items might not be well targeted to the most able persons. Otherwise, the item content matched the ADL abilities of the persons, and no further gaps influencing the responses of the persons were detected. The two gaps in the distribution that slightly exceeded 0.50 logit (0. 67, 0.75 logit) were located at the lowest performance categories among the easiest items, and no people were located across from those items.

While this evidence supports the validity of the A-ONE ADL scale, it is clear that there are also limitations. That is, while one can conclude that the items do represent ADL, such items are not hard enough for more able persons. This raises the question as to whether it would be appropriate to expand the content of the ADL scale to include harder items, or if it is more appropriate to recognize that that ADL scale might not be an appropriate evaluation instrument for testing more able persons.

Global NBI scales

When considered critically, the evidence based on content for the common short-form scale (Study III) was not strong. That is, this scale had poor targeting of items to persons, suggesting limited potential for differentiating between persons with lower levels (see Figure 3). Moreover, the large number of items omitted due to misfit likely resulted in the extensive and

large gaps detected in the hierarchy, underlining poor targeting. Another finding that failed to support validity based on content was the need to omit a large number of items that clinically are relevant to many of the people for whom the A-ONE is intended to be used, and which represent the construct of neurobehavioral limitations impacting ADL task performance. Thus, while a common global NBI scale that could be used across several diagnostic groups would be ideal, both clinically, as it simplifies procedures, and for research, as it provides opportunity to make comparisons across different diagnostic groups, from a content validity perspective, we were unable to accomplish this goal.

The evidence based on content for the NBI-CVA scale (Study IV), on the other hand, was better, as indicated by better targeting (see Figure 4). Only two gaps could be detected exceeding the 0.50 logit; these gaps were located at the mean level of neurobehavioral impact for persons. Items on the scale extended the full range of neurobehavioral impact on ADL task performance for the persons in the study, thus presenting with sufficient spread. Further, on the NBI-CVA scale, there are several more difficult items that might be useful in differentiating between persons that have an even greater degree of neurobehavioral impairments impacting ADL task performance than did our sample. The NBI-CVA scale also included the motor items which are considered clinically important for evaluating persons with either LCVA or RCVA; as noted below, these items were omitted from the NBI-LCVA scale. Thus, from a content validity perspective, we were able to construct a single global scale for use with persons diagnosed with hemispheric CVA that allows us to use one scale both in occupational therapy practice and research. Overall, as I will discuss below, the NBI-CVA scale seems to have a clinical advantage over the common short-form of the NBI scale for evaluating persons diagnosed with CVA. Obviously, further research is needed to confirm this assumption.

Diagnosis-specific scales

Targeting of items to persons revealed that there were several difficult items not matched by any persons on three of the four diagnosis-specific scales (LCVA, DAT, and DU), and there was a ceiling effect on the LCVA scale where nine persons were not matched by any items (Study III). There also were gaps at the easier end of all scales, where items needed for differentiating between more able persons were not available. The smallest gaps and best targeting was obtained with the RCVA scale, where all the motor items were retained.

Evidence for content validity also includes examination of content and ordering of the hierarchy of the items, and if they match expectations. Item hierarchies for different groups were also compared in Studies II and III, but only in a descriptive way and not statistically. Overall, the retained items in each of the diagnosis-specific NBI scales matched the expected pattern of impairments commonly seen among the different diagnostic groups (ASA, 2009b; Bartels, 2004; Caplan, 1993). Comparing the agreement between the empirical diagnosis-specific item difficulty hierarchies obtained in Study II, and the expected progression of emergence of underlying neurobehavioral impairments among different diagnostic groups (ASA, 2009b; Árnadóttir, 1990; Bartels, 2004; Caplan, 1993), also provided evidence for validity based on content. The fact that the item hierarchies varied among groups also is consistent with expectations as it is well recognized that different groups have different underlying neurobehavioral impairments.

More specifically, the results from Study II indicate that the types of neurobehavioral impairments seen among persons with RCVA matched the items retained on the NBI-RCVA diagnosis-specific scale. For the NBI-LCVA scale, however, the motor items, which are frequently detected in this diagnostic group, were unexpectedly omitted, resulting in a mismatch between the expected pattern of impairments and the items retained in the final scale. The motor items were also excluded from the DAT scale, but these impairments are seen among persons only in the later stages of the disease. Possible reasons for why these items were omitted are discussed below.

In conclusion, the content evidence for the four diagnoses-specific scales was acceptable, but not strong. It would be beneficial to add easy items on most of the scales.

Evidence based on response processes

One advantage of Rasch measurement methods is the ability to examine the validity of persons' response patterns to items by means of goodness-of-fit statistics for each person (Lim et al., 2009; Smith, 2001; Wolfe & Smith, 2007b). We did not, in any of our studies, address person response processes, or report any results related to the hierarchical ordering of people or goodness of fit of the persons to the respective Rasch models for the ADL scale or any of the NBI scales. While this may be viewed as a limitation, our focus has been on first attempting to develop valid scales; clearly there is a need to evaluate person response validity in future research.

Evidence based on internal structure

If the items are not dichotomous, the examination of evidence based on internal structure typically begins with consideration of the psychometric properties of the rating scale. In the case of the ADL scale, after the two categories were collapsed, the rating scale was shown to be psychometrically sound (Linacre, 2002), thus providing evidence for internal structure (Study I). All of the NBI scales were based on dichotomous items, so analysis of the rating scale was not appropriate.

When item goodness of fit is related to evidence to support internal structure, goodness of fit can be viewed as a support for, but not as confirmation of, unidimensionality. Unidimensionality is not confirmed unless PCA analyses and examination of DIF support a single Rasch factor (Lim et al., 2009; Smith, 2001; Wolfe & Smith, 2007b). Before I discuss these aspects of evidence for internal structure, it is important to consider that some authors (Goodwin, 2002; Lim et al., 2009) have suggested that factor analyses can be used to provide validity evidence based on internal structure. CTT studies of the A-ONE included exploratory factor analysis and revealed three factors for the ADL scale and two factors for the NB Specific scale (Árnadóttir, 1990); such results might indicate that construction of unidimensional ADL or NBI scales is not possible. Analysis of internal consistency of the ADL items (including the communication items) based on Cronbach's alpha (Steultjens, 1998; Árnadóttir, unpublished data) also yielded questionable results in terms of a CTT view of unidimensionality, (see Table 1).

Through the use of MTT, in contrast to the CTT results, we did obtain an evidence base to support internal structure of the ADL and most of the NBI scales. More specifically, after omitting the communication items, thought to represent a different construct (i.e., a second factor), the remaining self-care and mobility items fit the Rasch rating scale model for the ADL scale within the expected 5% type I error limit (Study I). Unidimensionality was further supported by acceptable PCA values where 85% of the variance was explained by the Rasch factor, and only 3.6% of the unexplained variance was accounted for by the first contrast. DIF analyses of the ADL scale were not performed, and clearly there is a need for such research in the future.

Evidence to support internal structure of the global NBI scales varied. That is, our attempt to create a common global scale (excluding motor items) in Study II failed, as despite acceptable goodness of fit of the items, the PCA analysis failed to support unidimensionality. Thus, we abandoned our attempt to create a common scale and instead created diagnosis-specific NBI

scales. In Study III we tried a different approach by trying to find a subset of items that fit each of four diagnosis-specific scales and then attempted to create a common short-form NBI scale (including motor items). While we obtained acceptable goodness of fit of items and acceptable PCA results, DIF was extensive. Combined with poor evidence to support content for this common short-form of the NBI scale, the extensive DIF suggests a threat to internal structure. DIF analyses were not performed for the NBI-CVA scale, but should be in the future.

Evidence to support internal structure based on item fit for diagnosis-specific NBI scales was realized, but only after removal of several misfitting items from each scale (Studies II and III). PCA analyses confirmed unidimensionality of four diagnosis-specific NBI scales (Study III).

The stringent item exclusion criteria applied for the Rasch analysis of the NBI scales should also be considered in this context. That is, we set our criteria for item fit such that we accepted no misfitting items (Studies II and III). As noted above, this resulted in the motor items being omitted from both the NBI-LCVA scale and the DAT scale (Study III); the motor items were not included in the analyses in Study II. Thus, using stricter criteria with regard to internal structure, validity evidence in relation to content was reduced. The lack of fit of the motor items to the Rasch model for the DAT scale may reflect the fact that few persons with very severe dementia were included in the study sample as there were no participants with DAT from long-term care facilities.

Evidence based on relation to other variables

When evaluating evidence in relation to other variables, it is important that scores or measures from one scale demonstrate the expected relation with scores or measures from the other scale. Scales that are thought to be related should have higher relationships and scales though to be related to different constructs should have lower relationships. Likewise, when a tool is used to evaluate for differences between groups, those thought to differ should have mean measures that differ significantly, and no differences should be found between groups thought to be similar (AERA, 1999). Previously, raw item scores had been compared with scores from other instruments evaluating the same construct (see Table 1). Thus, information from two scales used to evaluate the same construct, more specifically the A-ONE ADL items scores versus the Barthel Index scores were correlated, and the NBSIS scale scores were compared to the Mini Mental State Examination (MMSE) (Steultjens, 1998). Correlations of ADL and NB raw item scores had also been performed (Árnadóttir, 1990) (see Table 1). While these studies

have supported validity of the A-ONE scales in terms of relation to other variables, it was not possible to evaluate relations among total scores as it was recognized that summing ordinal raw scores was not valid.

Now, with the possibility to generate measures for the ADL scale and a NBI-CVA scale, it has become possible to evaluate the relationship between these two sets of measures. Because one scale was designed to evaluate ADL ability, and one was designed to evaluate the impact of neurobehavioral impairments on ADL task performance, we expected that the two tools might be related but that the relation should not be high — the two scales are used to measure different but overlapping constructs. As expected, we found a moderate correlation between the two scales, providing support for validity in terms of relation to other variables (Study IV).

Because we were able to construct an NBI-CVA scale to be used with persons with either RCVA or LCVA, we were also able to compare the impact of neurobehavioral impairments on ADL tasks performance between these two groups (Study IV). Earlier research has shown that these two groups do not differ in mean ADL ability (Bernspång & Fisher, 1995; Gardarsdóttir, & Kaplan, 2002; Rexroth, Fisher, Merritt, & Gliner, 2005; Shiotsuka, Burton, Pedretti, & Llorens, 1992). Nevertheless, we used ANCOVA procedures to control for any possible differences in mean ADL ability that might exist between the two samples included in our study. We found that the mean NBI measures also did not differ between groups, providing the first such study of this type.

Evidence based on consequences of testing

One consequence of testing is related to disruption of the measurement system such that test bias, in the form of DTF, occurs. That is, while DIF suggests that one group may be more likely to obtain higher scores than another, and this can be viewed as a threat to unidimensionality, and in turn, a threat to validity based on internal structure, the presence of DIF alone is not always viewed as a serious threat (Borsboom, 2006; Penfield & Camilli, 2007). In contrast, if DTF is present, there is a serious threat to validity in a form that can be considered a consequence of testing. In Study III, we did evaluate for DTF. The decision that DTF was not present was based on comparisons of person NBI measures (diagnosis-specific vs. common short-form scale) for each diagnostic group, where obtained correlations ranged from r = .90 to r = .99 and all participants except one client with RCVA presenting with NBI measures within the 95% confidence interval control lines.

Otherwise, evidence based on consequences was not addressed in the studies included in this thesis. Rather, validity related to consequences of testing was an important factor underlying the rationale for the studies included in this thesis. That is, the potential usefulness of measures generated from tests is a benefit, and thus one type of consequence, but the A-ONE scales have been limited in terms of their potential to be used as outcome measures. While clinicians have reported that they can use the A-ONE to gather important descriptive information that helps them to set goals and plan intervention, also an important source of validity related to consequences of testing, the A-ONE could not be used to evaluate the effectiveness of those interventions. Thus, one of the most important outcomes of this thesis is that the first step toward realizing that goal has been completed. Obviously, until evidence is gathered that the A-ONE scales are actually useful in evaluating change, there will remain no evidence that they are valid for such purposes.

New reliability evidence: From CTT to MTT

CTT studies of the A-ONE have supported interrater reliability (Árnadóttir, 1990, 2005), but otherwise, little evidence for the reliability of A-ONE raw scores was available. Now, with the ability to generate measures of overall ADL ability and neurobehavioral impact, it is important to consider further evidence for reliability based on MTT. The reliability evidence generated in the studies in this thesis included separation indices and related reliability coefficients, specifically the Rasch equivalent of Cronbach's alpha reported in the form of an *R* value. Another unique form of reliability available from Rasch analyses is a standard error (*SE*) for each person's measure. The *SE* is of particular value as it provides an index at the level of the individual that informs us of the potential for measuring change (Bond & Fox, 2007; Harvill, 1991).

Analyses of all the scales developed by the MTT studies revealed a person separation index of two or more distinct strata, indicating that the items reliably separate persons. These finding may also provide support that the scales are sensitive enough to be used to measure change in performance of the persons involved, which is a prerequisite for using them as an outcome measure. The more strata detected, the more sensitive the measure (Smith, 2001). Hence, in this case, acceptable separation reliability may support the long-term goal of creating valid measures that can be used to evaluate the effectiveness of intervention.

When considering separation, the higher the separation, the better. While separation indices above 1.80 are considered acceptable, values of 2.00 or higher are preferred. The ADL scale and the common NBI-CVA scale both had separation indices above 2.00 (Studies I and IV). Among the diagnosis-specific scales, only the NBI-RCVA and DU scale had separation indices above 2.00 (Study III). The associated reliability coefficients were all above $R = .70$ which also supports reliability of the measures.

The mean *SE* for person measures, recognized to vary along the length of the scale, ranged from 0.51 logit for the NBI-RCVA scale to 0.70 logit for the common short-form of the NBI scale. Since differences in measures ≥ 2 *SE* indicate a statistically significant difference between two measures (Harvill, 1991), it appears that rather large differences will be needed if the NBI scales are to be realized as sensitive measure of change. The mean *SE* for the ADL scale was 0.54 logit, but was not reported in Study I. In a similar manner, if ≥ 2 *SE* indicate a statistically significant difference between two measures, it seems probable that the ADL scale may become a sensitive measure of change. Of course, future research will be needed to verify or refute the usefulness of both the ADL and the NBI scales.

From idea to practice: Implications for practice

When I entered into a second cycle through the integrated profession model, and planned the implementation of a measurement cycle that became this thesis (see Figure 5), my decision and these studies were motivated by a clinical demand to generate evidence for the effectiveness of services within a climate where payment issues have become more and more pressing. New statements related to measurement principles were made in the form of Rasch model assertions and these were subsequently evaluated in different MTT research studies. To complete the measurement cycle, there remains the need to return the results of the constructed unidimensional scales of the A-ONE back into practice through publication and continuing education courses. We have now published or submitted for publication the results of these studies. But publication is not enough. Thus, it is time to consider how practice may best benefit from the results of this thesis. Which scales are clinically most useful, and how can they best be used so that they add to clinical usefulness of the original version of the A-ONE? What I will present is based on my own speculation, and thus, will need to be verified clinically and evaluated in future research.

Clinical and research use of the new A-ONE scales

If we turn back to the OTIPM presented briefly in the Introduction, the ADL scale of the A-ONE already can be used clinically to report observational information describing the quality of actions observed during ADL task performance and need for assistance. Information from the initial evaluation can also be used to form goals for intervention, and influence the content of the intervention, just as before the MTT studies. It is now potentially possible that ADL and NBI measures can be obtained through the use of conversion tables that are generated by Winsteps (Linacre 1991–2008), where the raw ordinal scale scores are converted to Rasch measures (see case sample below). These measures can be compared to measures obtained from a later evaluation in order to measure change in ADL task performance and/or impact of neurobehavioral impairments on that performance.

Additionally, in terms of intervention, the item hierarchies based on the MTT studies (ADL scale, NBI-CVA scale, and diagnosis-specific scales) might be useful for choosing training tasks of appropriate difficulty level for a person and for guiding intervention. Pursuing further the use of A-ONE NBI measures in intervention, the occupational therapist could, for example, use tasks that address certain errors (impairments), and the choice of which error(s) to address first during intervention could be based on the item's location on the hierarchy, perhaps ones close to performance level of the person, rather than ones that are far from the person's NBI measure as they may be harder to change through intervention.

All four of the diagnosis-specific NBI scales might be able to be used to measure change. The global NBI scales have the added potential of comparing measures from individuals within different diagnostic groups. Across-diagnostic-group comparisons could be an asset for program evaluation and research studies. Thus, studies like Study IV, where we examined for differences in impact of neurobehavioral impairments on ADL task performance in persons with RCVA and LCVA, can be performed. However, because of omission of many clinically important items from the common short-form scale, including motor items, motivation, and working memory, the NBI-CVA scale seems much more likely to be feasible for clinical use. Even the diagnosis-specific NBI-LCVA scale seems to be less useful clinically as the motor items that frequently impact ADL task performance of persons with LCVA had to be omitted.

Case sample

I will now demonstrate, using the case sample presented in the Rationale and Introduction, how the newly developed A-ONE scales might be used with a person diagnosed with LCVA. She is a person who needs variable physical assistance with all items in the four ADL domains (Dressing, Grooming and hygiene, Transfers and mobility, and Feeding) of the ADL scale of the A-ONE. Her obtained total raw ordinal score is 12. When I look up this figure in the conversion table developed using the Winsteps program, I learn that her ADL measure is -2.40 logits (*SE* = 0.42 logit).

Scoring the NB items reveals that her limitations in ADL task performance and resulting diminished independence are related to the impact of several neurobehavioral impairments including Motor apraxia, Ideational apraxia, Perseveration, Organization and sequencing problems, Right motor impairment, Sensory and Expressive aphasia (as indicated by scores on the NBSIS of the A-ONE) in addition to some impairments from the NBPIS. Her total raw score for the CVA-NBI scale is 27. When I look up this figure in the conversion table for the CVA-NBI scale, I find that her NBI measure is 0.19 logit (*SE* = 0.33 logit).

When I evaluated her again, after 4 months of occupational therapy, I find that she has made improvements in both ADL task performance and decreased frequency of neurobehavioral errors. More specifically, her ADL measures is now 1.64 logits (*SE* = 0.34 logit) based on a raw score total of 42. Comparison of her ADL measures from the initial evaluation (-2.40 logits, *SE* = 0.42 logit) to the follow up evaluation (1.64 logits, *SE* = 0.34 logit) reveals significant improvement (4.04 logits), as statistical significance is reflected by a change in ADL ability that is greater than the sum of pre- and post-*SE* values (0.42 + 0.34 = 0.76 logit) (Harvill, 1991). Similarly, comparison of her NBI-CVA measures for the initial evaluation (0.19 logit, *SE* = 0.33) and the follow up evaluation raw score of 16 (-1.15 logits, *SE* = 0.38 logit) again reveals a significant improvement (1.34 logits compared to her summed *SE*s = 0.71 logit).

The results of the studies included in this thesis, combined with consideration of how these results might be returned to practice, do indicate that the new ADL and some of the NBI scales of the A-ONE have successfully been Rasch analyzed. Development of conversion tables to convert the ordinal scores, recorded after observation of ADL task performance, to interval measures (both ADL and NBI) are under development.

The uniqueness of the findings of the present studies is the possibility to evaluate not only ADL ability and what types of neurobehavioral impairments impact ADL task performance, but also the magnitude of NBI

from the view of the frequency of impairments determined by observed errors impacting ADL task performance. This magnitude is not dependent on type of impairment and thus, the magnitude of impact on right and left CVA persons can be compared directly.

Methodological considerations

Participants

In relation to participants included in the study, lack of the most severely demented individuals and possibly the persons with very severe CVA, needs to be considered in relation to the obtained scales and use of the scales clinically. Not having these individuals in the analysis may have lead to exclusion of some clinically important items, such as the motor impairment items from several NBI scales.

Another methodological limitation related to the sample includes the fact that all of the participants were from Icelandic rehabilitation and geriatric wards. Thus, generalization of the results to performance of persons with other neurological conditions may be limited. Further, while there is no reason to suspect cultural specificity related to ADL task performance or the impact of neurobehavioral impairments on ADL task performance, the generalization of these findings to other world regions should be done cautiously until cross-cultural studies can be implemented.

Rating scales

In terms of the psychometric properties of the ADL rating scale, we used more stringent criteria than were recommended by Linacre (2002). That is, according to Linacre, it may not be necessary to collapse categories when threshold disordering is identified as disordering can occur merely because a category is rarely used. When too many categories result in raters not being able to differentiate between adjacent categories, collapsing categories is indicated; it is likely that scale reliability will improve (Linacre, 2002). In contrast, if the raters can differentiate between adjacent categories, collapsing categories can result in overall lower reliability and, in turn, lower sensitivity of the final measures for detecting change (Andrich, 1996; Stone & Wright, 1994). Thus, the need and impact of collapsing categories needs to be evaluated in more detail.

In terms of the NBI scales, dichotomizing of the rating scale items enhanced the potential to apply the newly constructed scales in practice using only simple conversion tables. Otherwise, we would have had to use a partial credit model to allow the inclusion of items with different rating

scales (Bond & Fox, 2007), and clinical application would then require the use of a computer program to accommodate for the different scales for different items, a procedure that would complicate clinical application by occupational therapists. Once the most useful NBI scale(s) have been identified, consideration of the pros and cons of expanding rating scale categories for all items can be addressed and evaluated through further research.

Misfit

For the Rasch analyses of the NBI scale, we maintained the same *MnSq* exclusion criterion of > 1.4 as we had used for the ADL scale. Our criterion was stricter than the recommended criterion of > 1.7 for analysis of clinical observations (Bond & Fox, 2007; Linacre, 1994). This may have resulted in unnecessary item exclusion. However, as the determination of unidimensionality of the NBI scales was based both on item goodness of fit and PCA, and PCA goes up when misfitting items are omitted, this may not have made a difference.

Ceiling effect and targeting of items to persons

When considering the ceiling effect on the ADL scale, one should keep in mind that the A-ONE is comprised of both an ADL scale and one or more potential NBI scales, and although some persons will have maximum scores on the ADL scale, they still may have had impairments detected by a NBI scale. Further, the ceiling effect of the NBI-LCVA scale seems to be related to omission of important items. As the NBI-CVA scale seems to work well for persons diagnosed with LCVA, it might be better to use that scale clinically for persons with LCVA.

The lack of items at higher levels of the ADL scale and some of the NBI scales could be related to what has been termed “lack of coverage” (Hudges, Dineen, Lai, & Cella, 2004). Resolution of this problem necessitates development of items that fill the void needed to allow differentiation of persons located at the level of the gap.

Clinical usefulness versus psychometric strength

The NBI scale studies indicate that both diagnosis-specific and common scales can be created, at least from a statistical perspective. However, the validity in terms of content, and, in turn, clinical usefulness, particularly of the common short-form scale, where many important items have been omitted, needs to be considered.

The emphasis on the importance of exploring DIF, as an important aspect of unidimensionality, has increased recently. In Study III, we developed several diagnosis-specific unidimensional scales, with different item hierarchies. Thus, not surprisingly, many items on the common short-form scale displayed with DIF. Conrad and associates (2007) suggested that different patterns of impairments could be related different diagnostic groups, which without doubt is the reason for the DIF in our study.

One solution when items demonstrate DIF is to split the items and form two or more new items, one for each of the different subgroups of persons (Tennant et al., 2004). When there is extensive DIF, such a solution would lead to very lengthy scales, and subsequently scales that likely would be too cumbersome for clinical use; computer scoring would become necessary. Thus, splitting items was not considered a suitable solution for resolving the DIF found among the NBI items of the A-ONE. Rather, the DIF on the common short-form scale was explored in the light of DTF. Results indicated that despite DIF, there is no DTF, or lack of fairness. Said in other words, we chose a clinically useful approach rather than one driven by statistical findings.

Context

It should be stressed that unidimensionality of performance errors detected during ecologically-relevant ADL task performance does not necessarily mean that unidimensionality of diverse neurobehavioral impairments would be obtained if they had been evaluated in a non-ecologically relevant context, where the focus is on presence and absence of impairments (as opposed to impact on ADL task performance). Moreover, errors observed during ADL task performance may not be detected based on neuropsychological evaluation, and neurobehavioral impairments detected during neuropsychological evaluation may not be observed to impact ADL. Finally, the errors detected using the NBI scales of the A-ONE are based on errors observed during ADL task performance, and cannot be generalized to performance of other daily life tasks without further study.

Software and statistical considerations

Computer programs change as new versions are developed, and so does expert opinion about statistical criteria. The newest version of Winsteps uses different methods to calculate PCA. This has lead to new criteria for acceptable PCA (Linacre, 1990-2009). The NBI scales developed in the Studies II and III, therefore, were based on the older criteria, and need to be reevaluated in light of newer criteria.

Further, my emphasis has changed with time. At the beginning of these studies, reporting goodness of fit, without so much emphasis on PCA and DIF was common. Subsequently, greater emphasis on PCA and DIF began to emerge in the literature. In part, emphasis appears to vary among different Rasch scholars, with some placing more emphasis on more rigid statistical criteria and with DIF analysis as a starting point; whereas others have placed more emphasis on goodness-of-fit, PCA, and clinical usefulness. In light of such changes and differences among Rasch scholars, it is comforting to recall Yerxa's (1967) comments to a student: "Don't worry about today's lesson, what you will learn in tomorrow's will make all of this obsolete" (pg. 156).

Recommendations for future research

The studies of this thesis have developed some new knowledge, and they have also generated many questions that need to be explored further. To continue the development of the A-ONE, the following areas could be studied further:

Items could be added on the ADL scale to diminish the ceiling effect. In that way, the ADL scale could be used with persons with higher levels of ADL ability, including those with only mild CVAs or in the early phases of dementia. Subsequently, new NBI items matching the new ADL items would need to be developed and evaluated through research. Further, addition of new NBI items needs to be considered where there is a ceiling effect, and also where there are gaps along the scales. If new NBI items are developed in relation to the development of new ADL items, this may address this need. If not, other solutions are needed. Another solution might be to try to develop items that differentiate more between different types of motor impairments (i.e., tremor, rigidity, weakness). For any studies where new items are included, collection of new data that includes both new and old items is needed so that the new data can be linked with previous data.

Studies on performance of individuals from other world regions and cultures need to be performed. To date, only studies of Icelandic and Dutch participants have been implemented; the Dutch results (Steultjens, 1998) were similar to the Icelandic results. Further, the participants of the studies included in this thesis were mainly limited to two diagnostic groups. Other groups with neurobehavioral impairments impacting ADL (e.g., persons with acquired head or spinal cord injuries) should also be added, and the validity and applicability of the ADL and NBI scales evaluated.

As noted earlier, the present studies focused on item structure more than person fit. Thus, future studies will need to explore evidence based on person response validity in more detail.

Conclusions

The following conclusions can be drawn from the results of the studies included in this thesis:

- Unidimensionality of the ADL scale can be obtained by removal of the two communication items and the possible future revision of the "Use knife" item. Rating scale structure can be improved by collapsing two adjacent scoring categories. Conversion tables can be made for therapists interested in using the A-ONE ADL scale as a linear outcome measure.
- This is the first study to establish theoretical hierarchies of diverse NBIs that have actually been observed in natural context of ADL task performance and that affect the quality of that task performance.
- Although NBIs are heterogeneous, they seem to belong to the same construct.
- The development of common NBI scales for use with different groups appears to be possible and clinically desirable. In contrast, the four diagnosis-specific scales are useful for generating measures, but limited to single groups of person, as opposed to having a potential for cross-diagnostic comparison.
- A collapsed scale, specifically the common short-form NBI scale which was developed in a more statistically-driven than applied manner, does not seem to be clinically valid and may have limited clinical use.
- In addition to being clinically useful in setting goals and guiding treatment decisions, the A-ONE scales now appear to have potential for use when measuring change. Thus, they likely can now be used to evaluate effectiveness of rehabilitation services and in quality assurance.
- While the studies in this thesis have generated evidence for validity and reliability of A-ONE measures, there remains a need for future research to continue the process of accumulating evidence to support the use of the A-One scale measures in research and practice.

Acknowledgements

I would like to acknowledge all my colleagues who contributed in different ways to the completion of this thesis. In particular I wish to thank:

My main supervisor *Anne G. Fisher*. Decades ago a supervisor of mine from the United States suggested that I contact Anne, as she would be the occupational therapist with the expertise required to assist with the psychometric development of the A-ONE. Everything has its time and place. The time is now, and the place is not the United States, but Umeå, Sweden. Drawing upon Anne's expertise in occupational therapy, psychometrics, and editorial skills has enabled me to more thoroughly develop the A-ONE. Anne's continuous support of this project has been invaluable.

Britta Löfgren, my co-supervisor, who ensured things moved along smoothly.

Birgitta Bernspång, Head of the Division of Occupational Therapy at Umeå University who welcomed me to the program and has provided me with important support throughout my studies.

Sigrún Garðarsdóttir, my supervisor at Landspítali University Hospital (LSH) in Iceland, who has never stopped believing in the A-ONE. My colleagues at LSH who have used the A-ONE in their work, including, but not limited to, *Edda Valtýsdóttir, Lillý H. Sverrisdóttir, Rósa Hauksdóttir, Sigríður Bjarnadóttir, Sigrún Ólafsdóttir*, and *Sigþrúður Loftsdóttir*.

My colleague and A-ONE trainer, *Eva Wæhrens* in Denmark, who encouraged me to take on the doctoral program in Umeå and accompanied me in the process. I must also acknowledge the other trainers who kept the A-ONE courses running internationally during my studies, including *Elsa van Schouwen, Esther Steultjens, Glen Gillen, Valerie Harris*, and *Sigrún*.

Valerie and *Sigrún*, also as teachers at the University of Akureyri, who taught part of my teaching load during the early stages of the studies, and other staff at the University who supported my studies.

Ingeborg Nilsson, my colleague within the doctoral student group, who led the way, and provided ongoing support and friendship. Other doctoral students, including *Michaela Munkholm, Maria Lindström*, *and Sólveig Ása Árnadóttir* are also acknowledged. All of my colleagues at the *Department of Community Medicine and Rehabilitation, Division of Occupational Therapy, Umeå University,* particularly within the research group, are also recognized. Finally, *Anders Kottorp,* for useful discussions and comments on manuscripts.

Ragnheiður Kristjánsdóttir, graphic designer for all her assistance.

Last, but not least, *my family*, for their unending patience and support.

_____

The first part of these studies was partially supported financially by grants from the University of Akureyri and Landspítali University Hospital in Iceland.

References

Alzheimers Association. Retrieved November 14, 2009 from http://www.alz.org/alzheimers_disease_what_is_alzheimers.asp

Alzheimers Society. Retrieved November 14, 2009 from http:// www.alzheimers.org.uk/

Alzheimer Society of Canada 1997-2009. Retrieved November 14, 2009 from http://www.alzheimer.ca/

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for education and psychological testing*. Washington, DC: American Educational Research Association.

American Heart Association (2009). *Statistical fact sheet – Populations: International cardiovascular disease statistics*. Retrieved November, 2009, from http://strokeassociation.org/downloadable/heart/123620401212INTL.pdf

American Occupational Therapy Association (2008). Occupational therapy practice framework: Domain and process, (2nd ed.). *American Journal of Occupational Therapy, 62*, 625–683.

American Stroke Association (2009a). *Types of stroke*. Retrieved November 14, 2009 from http://www.strokeassociation.org/ presenter.jhtml?identifier=1014

American Stroke Association (2009b). *How stroke affects the brain*. Retrieved November 14, 2009 from http://www. strokeassociation.org/presenter.jhtml?identifier=1052

American Stroke Association (2009c). *Impact of stroke*. Retrieved November 14, 2009 from http://www.strokeassociation.org/ presenter.jhtml?identifier=1033

Anastasi A. & Urbina S. (1997). *Psychological testing*, (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Andrich, D. (1996). Category ordering and their utility. *Rasch Measurement Transactions, 9,* 464–465. Retrieved March, 29, 2006, from http://www.rasch.org/rmt/rmt94f.htm

Asher, I. E. (Ed.). (2007). *Occupational therapy assessment tools: An annotated index* (3rd ed.). Bethesda, MD: American Occupational Therapy Association.

Árnadóttir, G. (1990). *The brain and behavior: Assessing cortical dysfunction through activities of daily living*. St. Louis, MO: Mosby.

Árnadóttir, G. (1999). Evaluation and intervention with complex perceptual impairment. In C. Unsworth (Ed.), *Cognitive and perceptual dysfunction: A clinical reasoning approach to evaluation and intervention* (pp. 393–454). Philadelphia, PA: F. A. Davis.

Árnadóttir, G. (2004a). Impact of neurobehavioral deficits on activities of daily living. In G. Gillen & A. Burkhardt (Eds.), *Stroke rehabilitation: A function-based approach*, (2nd ed., pp. 376–426). St. Louis, MO: Mosby.

Árnadóttir, G. (2004b). *Development versus dysfunction: Neurobehavioral perspective related to errors in occupational performance*. Poster session presented at the 7th European Congress of Occupational Therapy, Athens, Greece.

Árnadóttir, G. (2005). *Arnadottir OT-ADL Neurobehavioral Evaluation: Interrater reliability*. Poster session presented at the At-forum, Stockholm, Sweden.

Árnadóttir, G. (2008). Árangur af iðjuþjálfun einstaklinga með taugaeinkenni: Hentug ADL matstæki [Occupational therapy for persons with neurological disorders: Appropriate ADL outcome measures]. *Iðjuþjálfinn, 30*, 28–39.

Árnadóttir, G. (2009). *A-ONE training course: Lecture notes*. Unpublished manuscript.

Bartels, M. N. (2004). Pathophysiology and medical management of stroke. In G. Gillen & A. Burkhardt (Eds.), *Stroke rehabilitation a function-based approach* (2nd ed., pp. 1–30). St. Louis, MO: Mosby.

Benson, J. & Clark, F. (1982). A guide for instrument development and validation. *American Journal of Occupational Therapy*, 36, 789–800.

Bernspång, B., & Fisher, A.G. (1995). Differences between persons with right or left cerebral vascular accident on the Assessment of Motor and Process Skills. *Archives of Physical Medicine and Rehabilitation*, *76*, 1144–1151.

Bohlig, M., Fisher, W. P. Jr., Masters, G. N., & Bond, T. (1998). Content validity, construct validity and misfitting items. *Rasch Measurement Transactions*, *12*, 607. Retrieved July 8, 2004, from http:// www.rasch.org/rmt/rmt121f.htm

Bond, T.G. & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. (2nd ed). Mahwah, NJ: Erlbaum.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44(Supp*l. 3), S176–S181.

Bouwens, S. F. M., van Heugten, C. M., Aalten, P., Wolfs, C. A. G., Baarends, E. M., van Menxel, D. A. J. et al. (2008). Relationship between measures of dementia severity and observation of daily life functioning as measured with the Assessment of Motor and Process Skills (AMPS). *Dementia and Geriatric Cognitive Disorders, 25,* 81–87.

British Heart Foundation Statistics Website. Retrieved November 14, 2009 from http://www.heartstats.org/datapage. asp?id=8615 15

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society, 4,* 547–558.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport CT: Praeger.

Caplan, L. R. (1993). *Stroke: A clinical approach*. Boston: Butterworth-Heinemann.

Cella, D., & Chang, C-H. (2000). A discussion of item response theory and its applications in health status assessment. *Medical Care, 38(Supp*l. II), 66–72.

Chaytor, N, & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review, 13*, 181–197.

Chaytor, N., Schmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validitiy of executive functioning assessment. *Archives of Clinical Neuropsychology, 21,* 217–227.

Claesson, L., Lindén, T., Skoog, I., & Blomstrand, C. (2005). Cognitive impairment after stroke-impact on activities of daily living and costs of care for elderly people. *Cerebrovascular Diseases, 19*, 102–109.

Conrad, K. J., Dennis, M. L., Bezruczko, N., Funk, R. R., & Riley, B. B. (2007). Substance use disorder symtoms: Evidence of differential item functioning by age. *Journal of Applied Measurement, 8*, 373–387.

Cooke, D.M., McKenna, K., Fleming, J., Darnell, R. (2006). Construct and ecological validity of the Occupational Therapy Adult Perceptual Screening Test (OT-APST). *Scandinavian Journal of Occupational Therapy, 13*, 49–61.

Crepeau, E. B., Cohn, E. S., & Boyt Schell, B. A. (2003). Occupational therapy practice. In E. B. Crepeau, E. S. Cohn, & B. A. Boyt Schell (Eds.), *Willard & Spackman's occupational therapy* (10th ed., pp. 27—45). Philadelphia: Lippincott Williams & Wilkins.

Davies, P. L. & Gavin, W. J. (1999). Measurement issues in treatment effectiveness studies. *The American Journal of Occupational Therapy, 53*, 363–372.

Davis, J., Craik, J., & Polatajko, H. J. (2007). Using the Canadian Process Practice Framework: Amplifying the process. In E. A. Townsend & H. J. Polatajko (Eds.), *Enabling occupation II: Advancing an occupational therapy vision for health, well-being, & justice through occupation,* (*pp.* 247–272). Ottawa, Ontario: Canadian Association of Occupational Therapists.

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(Suppl. 3), S50–S59.

Dickoff, J., James, P., & Wiedenbach, E. (1968). Theory in a prctice discipline. Part I: Practice oriented theory. *Nursing Research, 17, 415–435*.

Doble, S. E., Fisk, J. D., & Rockwood, K. (1999). Assessing the ADL functioning of persons with Alzheimer's disease: Comparison of family informant's ratings and performance-based assessment findings. *International Psychogeriatrics, 11*, 1999, 399–409.

Donkervoort, M., Dekker, J., & Deelman, B.G. (2002). Sensitivity of different ADL measures to apraxia and motor impairments. *Clinical Rehabilitation, 16*, 299–305.

Edmans J.A., & Lincoln, N.B. (1990). The relation between perceptual deficits after stroke and independence in activities of daily living. *British Journal of Occupational Therapy, 53*, 139–142.

Fänge, A., Lanke, J., & Iwarsson, S. (2004). Statistical assessment of changes in ADL dependence: Three-graded versus dichotomised scaling. *International Journal of Rehabilitation Research, 27*, 305–309.

Fisher, W. P. (1993). Measurement-related problems in functional assessment. *American Journal of Occupational Therapy, 47*, 331–338.

Fisher, A. G. (1993). The Assessment of IADL motor and process skills: An application of many faceted Rasch analysis. *American Journal of Occupational Therapy, 47,* 319–329.

Fisher, A. G. (1998). Uniting practice and theory in an occupational framework: 1998 Eleanor Clarke Slagle Lecture. *American Journal of Occupational Therapy, 52*, 509–521.

Fisher, A. (2006a). Overview of performance skills and client factors. In H. M. Pendleton & W. Schulz-Krohn (Eds.), *Pedretti's occupational therapy*: *Practice skills for physical dysfunction* (6th ed., pp. 372–402). St. Louis, MO: Mosby Elsevier.

Fisher, A. G. (2006b). *Assessment of Motor and Process Skills. Vol. 1: Development, standardization, and administration manual* (6th ed.). Fort Collins, CO: Three Star Press.

Fisher, A. G. (2009). *Occupational Therapy Intervention Process Model: A model for planning and implementing top-down, client-centered, and occupation-based interventions*. Fort Collins, CO: Three Star Press.

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heineman, A. W., et al. (1994). Applications of conjoint measurement to the development of functional measures. *International Journal of Educational Research, 21*, 579–593.

Gardarsdóttir, S., & Kaplan, S. (2002). Validity of the Árnadóttir OT-ADL Neurobehavioral Evaluation: Performance in activities of daily living and neurobehavioral impairments in persons with left and right hemisphere damage. *American Journal of Occupational Therapy, 56*, 499–508.

Gauthier, L., & Gauthier, S. (1990). Assessment of functional changes in Alzheimer's disease. *Neuroepidemiology, 9,* 183–188.

Geyh, S., Cieza, A., Schouten, J., Dickson, H., Frommelt, P., Omar, Z., et al. (2004). ICF core sets for stroke. *Journal of Rehabilitation Medicine*, 44 (Suppl.), 135–141.

Geyh, S., Kurt, T., Brockow, T., Cieza, A., Ewert, T., Omar, Z. et al. (2004). Identifying the concepts contained in the outcome measures of clinical trials on stroke using the International Classification of Functioning, Disability and Health as a reference. *Journal of Rehabilitation Medicine*, 44,(Suppl.) 56–62.

Gillen, G. (2006). Cerebrovascular accident/stroke. In H. M. Pendleton & W. Schulz-Krohn (Eds.), *Pedretti's occupational therapy*: *Practice skills for physical dysfunction* (6th ed., pp. 802–838). St. Louis, MO: Mosby Elsevier.

Gillen, G. (2009). *Cognitive and perceptual dysfunction: Optimizing function*. St. Louis, MO: Mosby Elsevier.

Glymour, M. M., Berkman, L. F., Ertel, K. A., Fay, M. E., Glass, T. A., & Furie K. L. (2007). Lesion characteristics, NIH stroke scale, and functional recovery after stroke. *American Journal of Physical Medicine and Rehabilitation, 2007, 86*, 725–733.

Golden, C. J., Sawicki, R. F., & Franzen, M. D. (1990). Test construction. In G. Goldstein & M. Hersen (Eds.). *Handbook of psychological assessment* (2nd ed., pp. 21–40). New York: Pergamon Press.

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new standards. *Journal of Nursing Education, 41*, 100–106.

Goto, A., Okuda, S., Ito, S., Matsuoka, Y., Ito, E., Takahashi, A. et al. (2009). Locomotion outcome in hemiplegic patients with middle cerebral artery infarction: the difference between right- and left-sided lesions. *Journal of Stroke and Cerebrovascular Diseases, 18*, 60–67.

Granger, C. V., Hamilton, B. B., & Fiedler, R. C. (1992). Discharge outcome after stroke rehabilitation. *Stroke, 23*, 978–982.

Haertel, E. H. (2006). Reliability. In R. L. Brennan, (Ed.), *Educational measurement*, (4th ed.). Westport, CT: Praeger.

Hagedorn, R. (2001). *Foundations for practice in occupational therapy* (3rd ed.). Edinburgh: Churchill Livingstone.

Haigh R, Tennant A, Biering-Sørensen F, Grimby G, Marinček Č, Phillips S, et al. (2001). The use of outcome measures in physical medicine and rehabilitation within Europe. *Journal of Rehabilitation Medicine*, 33, 273–278.

Hammond, K. R. (1998). Ecological validity: Then and now. Retrieved October, 23, 2006 from http://brunswik.org/notes/essay2.html

Harvill, L. M. (1991). NCME instructional module: Standard error of measurement. *Educational Measurement: Issues and Practice, 10,* 33-41.

Haymes S.A., Johnston, A.W., Heyes, A.D. (2001). The development of the Melbourne Low-Vision ADL Index: A measure of vision disability. *Investigative Opthalmology and Visual Science,* 42, 1215–1225.

Heart and stroke Foundation of Canada. Help erase the effects of stroke. Retrieved November 14, 2009 from http://www.heartandstroke.com/ site/c.ikIQLcMWJtE/b.3483991/k.34A8/ Statistics.htm

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., & Granger, C. (1993). Relationships between impairment and physical disability as measured by the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 74,* 556–573.

Hudgens, S., Dineen, K., Webster, K., Lai, J-S, & Cella, D. (2004). Assessing statistically and clinically meaningful construct deficiency/ saturation: Recommended criteria for content coverage and item writing. *Rasch Measurement Transactions, 17,* 954–955. Retrieved August, 23, 2008, from http://www.rasch.org/rmt/rmt174d.htm

Johnston, M. V., Findley, T. W., DeLuca, J., & Katz, T. T. (1991). Research in physical medicine and rehabilitation: XII. Measurement tools with application to brain injury. *American Journal of Physical Medicine and Rehabilitation, 70*(Suppl. 1), S114–S129.

Johnstone, B., & Frank. R. G. (1995). Neuropsychological assessments in rehabilitation: Current limitations and applications. *Neuro Rehabilitation, 5,* 75–86.

Kielhofner, G. (2009). *Conceptual foundations of occupational therapy practice* (4th. Ed.). Philadelphia: F. A. Davis.

Kielhofner, G., & Forsyth, K. (2008). Therapeutic reasoning: Planning, implementing, and evaluating the outcomes of therapy. In G. Kielhofner, (Ed.), *Model of Human Occupation: Theory and application* (4th ed., pp. 143–154). Baltimore, MD: Lippincott Williams & Wilkins.

Korpelainen, J.T., Niilekselä, E., & Myllylä, V.V. (1997). The Sunnaas Index of Activities of Daily Living: Responsiveness and concurrent validity in stroke. *Scandinavian Journal of Occupational Therapy, 4*, 31–36.

Langhorne, P., Williams, B. O., Gilchrist, W., & Howie, K. (1993). Do stroke units safe lives. *Lancet, 342,* 395–398.

Law, M., Baum, C, & Dunn, W. (Eds.). (2005). *Measuring occupational performance: Supporting best practice in occupational therapy* (2nd ed.). Thorofare, NJ: SLACK.

Lim, S. M., Rodger, S., & Brown, T. (2009). Using Rasch analysis to establish the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation, 15*, 251–260.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions, 9*, 422–423. Retrieved March, 29, 2006, from http://www.rasch.org/rmt/rmt92b.htm

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2,* 266–283.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement,* 3, 85–106.

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer software. Chicago: Winsteps.com.

Linacre, J. M. (1991-2006). *A user's guide to Winsteps® Ministep Rasch-model computer programs.* Retrieved January, 10, 2007, from http://www.winsteps.com/aftp/winsteps.pdf

Linacre, J.M. (2008). WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com.

Linacre, J. M. (1991-2009). *A user's guide to Winsteps® Ministep Rasch-model computer programs.* Retrieved January, 15, 2010, from http://www.winsteps.com/aftp/winsteps.pdf

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. W., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75,* 127–132.

Llorens, L. A. (1986). Activity analysis: agreement among factors in a sensory processing model. *American Journal of Occupational Therapy, 40*, 103–110.

Lopez, W. (1996). Communication, validity and rating scales. *Rasch Measurement Transactions*, *10*, 482–483. Retrieved July 30, 2004, from http://www.rasch.org/rmt/rmt101k.htm

Mattingly, C., & Fleming, M. H. (1994). *Clinical reasoning: Forms of inquiry in a therapeutic practice*. Philadelphia, PA: F. A. Davis.

McAllister S. (2008). Introduction to the use of Rasch analysis to assess patient performance. *International Journal of Therapy and Rehabilitation, 15*, 482–490.

Medin, J., Nordlund, A., Ekberg, K. (2004) Increasing stroke incidence in Sweden between 1989 and 2000 among persons aged 30 to 65 years: Evidence from the Swedish Hospital Discharge Register. *Stroke, 35,* 1047–1051.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, 70, 308–312.

de Morton, N. A., Keating, J. L. , & Davidson, M. (2008). Rasch analysis of the Barthel Index in the assessment of hospitalized older patients after admission for an acute medical condition. *Archives of Physical Medicine and Rehabilitation, 89*, 641–647.

National Institute of Neurological Disorders & Stroke (2010). Stroke rehabilitation information. Retrieved February 1, 2010, from http://www.ninds.nih.gov/disorders/stroke/stroke_rehabilitation.htm

Neistadt, M. E. (1992). The Rabideau Kitchen Evaluation–Revised: An assessment of meal preparation skill. *Occupational Therapy Journal of Research, 12*, 242–255.

Neistadt, M.E. (2000). *Occupational therapy evaluation for adults*. Philadelphia: Lippincott Williams & Wilkins.

Nilson, Å. L., Sunnerhagen, K. S., & Grimby, G. (2005). Scoring alternatives for FIM in neurological disorders applying Rasch analysis. *Acta Neurologica Scandinavica*, *111*, 264–273.

Nuwer, M. R., Árnadóttir, G., Martin, N. A., Ahn, S. S., & Carlson, L. (1994). A comparison of quantitative electroencephalography, computed tomography, and behavioral evaluations to localize impairment in patients with stroke and transient ischemic attacks. *Journal of Neuroimaging, 4*, 82–84.

Nygård, L., Amberla, K., Bernspång, B., Almkvist, O., & Winblad,B. (1998). The relationship between cognition and daily activities in cases of mild Alzheimer's disease. *Scandinavian Journal of Occupational Therapy, 5,* 160–166.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). New York: Elsevier.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 6*. Retrieved January 12, 2008, from http://journals.tc-library.org/templates/about /editable/ pdf/ Perrone%20Forum.pdf

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960).

Rexroth, P., Fisher, A.G., Merritt, B.K., & Gliner, J. (2005). ADL differences in individuals with unilateral hemispheric stroke. *Canadian Journal of Occupational Therapy*, 72, 212–221.

Rijken, P. M. & Dekker, J. (1998). Clinical experience of rehabilitation therapists with chronic diseases: A quantitative approach. *Clinical Rehabilitation, 12,* 143–150.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1996). The structure of normal human attention: *Journal of International Neuropsychological Society, 2*, 525–534.

Salter, K., Jutai, J. W., Teasell, R., Foley, N. C., Bitensky, J., & Bayley, M. (2005). Issues for selection of outcome measures in stroke rehabilitation: ICF participation. *Disability and Rehabilitation, 27*, 507–528.

Schultz, S., & Schkade, J. K. (2003). Occupational Adaptation. In E. B. Crepeau, E. S. Cohn, & B. A. Boyt Schell (Eds.), *Willard & Spackman's occupational therapy* (10th ed., pp. 220—223). Philadelphia: Lippincott Williams & Wilkins.

Schultz-Krohn, W., & Pendleton, H. M. (2006). Application of the Occupational Therapy Practice Framework to physical dysfunction. In H. M Pendleton, & W. Schultz-Krohn, (Eds.), *Pedretti's occupational therapy practices Skills for physical dysfunction* (6th ed., pp. 28–52). St. Louis, MO: Mosby Elsevier.

Schwartz, M. F., Mayer, N. H., FitzpatrickDeSalme, E. J., & Montgomery, M. W. (1993). Cognitive theory and the study of everyday action disorders after brain damage. *Journal of Head Trauma Rehabilitation, 8,* 59–72.

Schwartz, M. F., Segal, M., Veramonti, T., Ferraro, M., & Buxbaum, L. J. (2002). The Naturalistic Action Test: A standardized assessment for everyday action impairment. *Neuropsychological Rehabilitation, 12,* 311–339.

Semkovska, M., Bédard, M-A., Godbout, L., Limoge, F., & Stip, E. (2004). Assessment of executive dysfunction during activities of daily living in schizophrenia. *Schizophrenia Research, 69,* 289–300.

Shiotsuka, W., Burton, G.U., Pedretti, L.W., & Llorens, L.A. (1992). An examination of performance scores on activities of daily living between elders with right and left cerebrovascular accident. *Physical and Occupational Therapy Geriatrics*, 10, 47–57.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51,* 541–565.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199-218.

Smith, Jr., E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281–311.

Spooner, D. M., & Pachana; N. A. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology, 21*, 327–337.

Steultjens, E. M. J. (1998). A-ONE: De Nederlandse Versie [A-ONE: The Dutch version]. *Nederlands Tijdschrift for Ergotherapie, 26*, 100-104.

Steultjens, E. M. J., Dekker, J., Bouter, L. M., van de Nes, J. C. M., Cup, E. H. C., & van den Ende, C. H. M. (2003). Occupational therapy for stroke patients: A systematic review. *Stroke, 34,* 676–687.

Stone, M.H., & Wright, B. D. (1994). Maximizing rating scale information. *Rasch Measurement Transactions, 8*, 386. Retrieved March, 29, 2006, from http://www.rasch.org/rmt/rmt83r.htm

Sundberg, G., Bagust, A., & Terént, A. (2003). A model for costs of stroke services. *Health Policy, 63,* 81–94.

Sveen, U., Bautz-Holter, E., Sødring, K.M., Wyller, T.B., Laake, K. (1999). Association between impairments, self-care ability and social activities 1 year after stroke. *Disability Rehabilitation*, *21*, 372–377.

Taylor, Jr., D. H., Schenkman, J., Zhou, J., & Sloan, A. S. (2001). The relative effect of Alzheimer's disease and related dementias, disability, and comorbidities on cost of care for elderly persons. *Journal of Gerontology: Social Sciences, 56B,* S285–S293.

Tennant, A. (2004). Disordered thresholds: An example from the Functional Independence Measure. *Rasch Measurement Transactions, 17*, 945–948. Retrieved July 30, 2004, from http:// www.rasch.org/rmt/rmt174a.htm

Tennant, A., Geddes, J. M. L., & Chamberlain, M. A. (1996). The Barthel Index: An ordinal score or interval level measure? *Clinical Rehabilitation, 10,* 301-308.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions, 20*, 1082-1084. Retrieved February, 4, 2008, from http://www.rasch.org/rmt/rmt204d.htm

Tennant A, Penta M, Tesio L, Grimby, G., Thonnard, J-L, Slade, A. et al. (2004). Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR Project. *Medical Care*, *42*(Suppl. 1) I-37—I-48.

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine, 35*, 105-115.

Tesio, L., Simone, A. & Bernardinello, M. (2007). Rehabilitation and outcome measurement: Where is Rasch-analysis going? *Europa Medicophysica, 43,* 417–426.

Tham, K., Bernspång, B., & Fisher, A. G. (1999). Development of the Assessment of Awareness of Disability. *Scandinavian Journal of Occupational Therapy, 6,*184–190.

Titus, M. N. D., Gall, N. G., Yerxa, E. J., Roberson, T. A., & Mack, W. (1991). Correlation of perceptual performance and activities of daily living in stroke patients. *American Journal of Occupational Therapy, 45*, 410–418.

Trombly Latham, C. A. (2008a). Conceptual foundations for practice. In M. V. Radomski, & C. A. Trombly Latham (Eds), *Occupational therapy for physical dysfunction* (6th, ed., pp. 1–20). Philadelphia: Lippincott Williams & Wilkins.

Trombly Latham, C. A. (2008b). Occupation as therapy: Selection, gradation, analysis and adaptation. In M. V. Radomski, & C. A. Trombly Latham (Eds.), *Occupational therapy for physical dysfunction* (6th, ed., pp. 358–381) Philadelphia: Lippincott Williams & Wilkins.

Unsworth, C. (1999). Reflections on the process of therapy in cognitive and perceptual dysfunction. In C. Unsworth, *Cognitive and perceptual dysfunction: A clinical reasoning approach to evaluation and intervention* (pp. 75–124). Philadelphia: F. A. Davis.

Wilson, B. A. (2002). Cognitive rehabilitation in the 21rst century. *Neurorehabilitation & Neural Repair, 16,* 207–210.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E. W., & Smith, Jr., E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement, 8*, 97-123.

Wolfe, E. W., & Smith, Jr., E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement, 8*, 204–234.

Wright, B. D. (1995). Diagnosing person misfit. *Rasch Measurement Transactions, 9*, 430-431. Retrieved March, 29, 2006, from http://www.rasch.org/rmt/rmt92h.htm

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70,* 857–860.

Wright, B. D., & Linacre, J. M., (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370. Retrieved October, 19, 2003, from http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Yavuzer, G., Küçükdeveci, A., Arasil, T., & Elhan, A. (2001). Rehabilitation of stroke patients: Clinical profile and functional outcome. *American Journal of Physical Medicine and Rehabilitation, 80*, 250–255.

Yerxa, E. J. (1967). The Eleanor Clarke Slagle Lecture: Authentic Occupational Therapy. *American Journal of Occupational Therapy, 21*, 155–173.

Yerxa, E. J. (1994). In search of good ideas for occupational therapy. *Scandinavian Journal of Occupational Therapy, 1,* 7–15.