From the Department of Woman and Child Health
Karolinska Institutet,
Stockholm, Sweden

The Assisting Hand Assessment: continued development, psychometrics and longitudinal use

Marie Holmefur

Stockholm 2009
To Niklas,
Jonathan
and Esther
ABSTRACT

For most people, using both hands together is a natural part of everyday activities. But for children with a unilateral disability in the arm and hand, activities requiring two hands are often a challenge. In order to evaluate interventions that optimise the use of the affected or assisting hand, valid and reliable measures are needed. The Assisting Hand Assessment (AHA) is a newly developed test with the expressed purpose of measuring and describing how effectively children with a unilateral disability in arm and hand use their assisting hand during bimanual tasks. An AHA assessment involves a video-taped play session, which elicits spontaneous use of the hands by using toys that require bimanual handling. The play session is scored using a 4-point rating scale for 22 items which are describing object-related hand actions. The AHA was initially validated for children aged 18 months to 5 years with unilateral cerebral palsy (CP) or obstetric brachial plexus palsy (OBPP) (Small Kids AHA). A test kit for older children containing two board games suitable for children aged 6-12 years has been developed (School Kids AHA). The aim of this thesis was to evaluate the psychometric properties of the AHA including the expanded age group and to describe the longitudinal development of assisting hand use in children with unilateral CP.

The validity of the AHA for ages ranging from 18 months to 12 years was investigated. Using a Rasch measurement model, 409 AHA’s of children with unilateral CP or OBPP were analyzed. Results showed that the items measure a unidimensional construct and that the four-point rating scale overall functions well. An excellent targeting between the item difficulties and the children's abilities was found. A high person separation indicates that the AHA can effectively separate between children of different ability into seven distinct ability strata. Differential item functioning between diagnostic groups was identified in some items and discussed.

The reliability of the AHA as regards the inter- and intrarater, retest and alternate forms reliability was investigated. In all trials high reliability coefficients (Intraclass Correlation Coefficients) were found, ranging from 0.97 to 0.99. The retest evaluation indicated that the smallest detectable difference measurable with the AHA is a change of 4 points (6% of the scale). It was further found that all tested versions, the Small Kids vs. School Kids AHA and the two board games in School Kids AHA, produce equal results and measure the same construct.

The development of assisting hand use between the ages 18 months and 8 years was investigated among 43 children with unilateral CP. The children were regularly assessed using the AHA for on average 4.5 years. All children increased their assisting hand use during the course of the study. A non-linear mixed model was used to create separate average development curves for children at different levels on the Manual Ability Classification System (MACS). The maximum level (limit) of development differed between children in MACS levels I-III. The rate of change was similar in levels I and II and significantly slower in children at level III. The children were also divided into two groups based on their AHA-score at 18 months. The limit and rate of change differed significantly between the groups. The AHA score at 18 months can be used for approximate prediction of future development of assisting hand use.

Keywords: Assisting Hand Assessment, measurement, hand function, children, unilateral cerebral palsy, hemiplegia, longitudinal, psychometrics, Rasch analysis
LIST OF PUBLICATIONS

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


### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHA</td>
<td>Assisting Hand Assessment</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>CAOT</td>
<td>Canadian Association of Occupational Therapists</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CP</td>
<td>Cerebral Palsy</td>
</tr>
<tr>
<td>df</td>
<td>Degrees of freedom</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>GA</td>
<td>Gestational Age</td>
</tr>
<tr>
<td>GMFCS</td>
<td>Gross Motor Function Classification System</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass correlation coefficient</td>
</tr>
<tr>
<td>MACS</td>
<td>Manual Ability Classification System</td>
</tr>
<tr>
<td>MnSq</td>
<td>Mean Square</td>
</tr>
<tr>
<td>OBPP</td>
<td>Obstetric Brachial Plexus Palsy</td>
</tr>
<tr>
<td>OT</td>
<td>Occupational Therapist</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Components Analysis</td>
</tr>
<tr>
<td>PVL</td>
<td>Periventricular lesion</td>
</tr>
<tr>
<td>QUEST</td>
<td>Quality of Upper Extremity Skills Test</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SDD</td>
<td>Smallest Detectable Difference</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard Error of Measurement</td>
</tr>
<tr>
<td>SS</td>
<td>Sum of Squares</td>
</tr>
</tbody>
</table>
INTRODUCTION

The hands are essential for our performance of everyday activities. For most people, activities are automatically carried out with both hands cooperating in a smooth and coordinated manner. This bimanual performance is so common and habitual that it is only noticed when there is a disturbance, e.g., a cut in the finger or when carrying a baby while doing other things. For children with a disability in one hand or arm, the situation is different. Tasks that can be performed single-handedly are preferably carried out with the unaffected, well-functioning hand. When both hands are needed tasks become much more complicated. Most commonly, the affected hand will hold and stabilize an object while the unaffected hand is manipulating it. The success of these tasks depends on the ability to use the affected hand as an effective help, or assisting hand (Krumlinde-Sundholm & Eliasson 2003). Effectiveness of assisting hand use is therefore an important focus for measurement of and interventions for these children. A valid and reliable measure with this focus is needed as a tool in both clinical practice and research.

The Assisting Hand Assessment (AHA) is a newly developed test with the purpose of measuring and describing how effectively children with a unilateral disability use their affected hand in bimanual tasks (Krumlinde-Sundholm & Eliasson 2003). In this thesis the psychometric properties of the AHA will be explored and the AHA will be used to describe development of assisting hand use.

MEASUREMENT OF ASSISTING HAND USE

An Occupational Therapy perspective

Measurement of hand function in children is often carried out by the occupational therapist (OT), though, the focus of the OT when meeting a family with a child with a unilateral disability in the arm and hand is however much broader than just hand function. The OT’s domain of concern is the entire occupational performance of the child (Law & Baum 2005; Stewart 1996). The OT sees hand function in terms of the impact it has on the occupations the child wishes, needs or is expected to perform in his or her environment (Law, King et al. 2005). There are a number of practice models within the field of occupational therapy that explore aspects of occupational performance. One of these models is the person-environment-occupation model, in which three circles are used to illustrate the interrelationship between the parts of the model, which includes aspects of the person, the occupation and the environment (Figure 1). Occupational performance is the outcome of the transaction between these three (Law et al. 1996).

The OT also has a family-centred perspective, which means that only occupations crucial to the child and family will be the focus of measurement and interventions (King, Teplicky et al. 2004). This also means that measurement should start with finding the occupational performance issues that are of concern to the child and family. When these issues have been identified the next step is to search for reasons for the limited occupational performance. Reasons may be found in the person, occupation or...
environment, or in the interaction between these (Law et al. 2005). When looking for reasons for limited occupational performance within the person, performance components such as hand function may be assessed. There are a number of different measures or tests available for measuring hand function in the child. The choice of measure should be guided by the overall occupational performance issue identified and by the purpose of the test (Law et al. 2005; Rosenbaum 1998).

The Assisting Hand Assessment (AHA)

The Assisting Hand Assessment (AHA) is an observation test designed to measure and describe how effectively children with unilateral disabilities in arm and hand use their affected hand in bimanual tasks.

Characteristics of the AHA

The AHA is a criterion-referenced test that was developed and validated specifically for children with unilateral disabilities in the arm and hand, more specifically unilateral cerebral palsy (CP) and obstetric brachial plexus palsy (OBPP). The AHA measures the spontaneous use of the assisting hand and not the child’s maximum capacity. Furthermore, only observable performance is assessed, and not the underlying impairment.

Development

The AHA was developed by Lena Krumlinde-Sundholm and Ann-Christin Eliasson at the Department of Woman and Child Health, Karolinska Institutet. They found that there was no test available with the specific purpose of measuring how children with unilateral disabilities use their affected hand in bimanual tasks. They developed a test kit by selecting toys that would elicit spontaneous use of both hands. These toys were gathered in a children’s suitcase, called the AHA suitcase (Figure 2). They videoed play sessions during which children were encouraged to engage in spontaneous play in interaction with a therapist. From observation of the videos they described object-related hand actions in the affected hand and in the collaboration of both hands seen while children played. These hand actions were all considered to be important for the
effective use of the assisting hand and were reformulated as items. A rating scale was devised and the performance required for each scale step was defined. This resulted in a manual with scoring criteria for all items and scale steps.

The assessment

An AHA assessment is conducted in two stages. In the first stage a play session is recorded on video. In the play session the child and the therapists sit opposite each other at a table and the camera is placed behind the therapist. The child and the therapist play with the toys together, the focus being on an enjoyable session. While playing the therapist hands the toys to the child one at the time, and lets the child try all the toys in the suitcase during the 10- to 15-minute session. There is no right or wrong way to handle the toys, so the child can do what he or she likes, and if needs be the therapist should instruct or demonstrate to avoid frustration and keep the session enjoyable.

In the second stage the child’s way of using the assisting hand is observed on the video and the performance is scored on 22 items on a four-point rating scale. The items in the AHA scale are listed in Table I. The scale steps are ordered and represent quality of observable performance. The four steps on the rating scale are: 4 (efficient), 3 (somewhat efficient), 2 (ineffective) and 1 (does not do). The items and criteria for each scale step and item are thoroughly described in the AHA Manual (Krumlinde-Sundholm, Holmefur et al. 2007).

The scores

The raw scores gained can be added to a sum score (range 22-88 points). They can also be converted to scaled scores (range 0-100). They both constitute ordinal-level data. logits which are interval-level data are obtained from a Rasch analysis. The range of the AHA-scale is -10.18 to 8.70 logits.

Use of the AHA

The AHA is scored on a software-based score sheet. On the first page items are arranged in logical order for ease of scoring. In the second page, which is used after scoring for planning of intervention, items are arranged in a hierarchical order, obtained using Rasch analysis, with the easier items first. On this page items with scores of 3 or 4 are marked green, indicating the strengths of the child’s performance. Items with scores of 1 or 2 are marked red, indicating ineffective performance or that actions have not been performed. Special attention should be paid to the items in the area where scores shift from green to red. These are probably the items that are closest to the child’s current ability level and the easiest for the child to achieve next. This
Introduction

information can guide formulation of goals for intervention (Krumlinde-Sundholm et al. 2007).

<table>
<thead>
<tr>
<th>Table I. Items in the Assisting Hand Assessment</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>General use items</td>
<td>Approaches objects</td>
</tr>
<tr>
<td></td>
<td>Initiates use</td>
</tr>
<tr>
<td></td>
<td>Chooses assisting hand when closer to objects</td>
</tr>
<tr>
<td>Arm use items</td>
<td>Stabilizes by weight/support</td>
</tr>
<tr>
<td></td>
<td>Reaches</td>
</tr>
<tr>
<td></td>
<td>Moves upper arm</td>
</tr>
<tr>
<td></td>
<td>Moves forearm</td>
</tr>
<tr>
<td>Grasp-release items</td>
<td>Grasps</td>
</tr>
<tr>
<td></td>
<td>Holds</td>
</tr>
<tr>
<td></td>
<td>Stabilizes by grip</td>
</tr>
<tr>
<td></td>
<td>Readjusts grip</td>
</tr>
<tr>
<td></td>
<td>Varies type of grasp</td>
</tr>
<tr>
<td></td>
<td>Releases</td>
</tr>
<tr>
<td></td>
<td>Puts down</td>
</tr>
<tr>
<td>Fine motor adjustment items</td>
<td>Moves fingers</td>
</tr>
<tr>
<td></td>
<td>Calibrates</td>
</tr>
<tr>
<td></td>
<td>Manipulates</td>
</tr>
<tr>
<td>Coordination items</td>
<td>Coordinates hands/arms</td>
</tr>
<tr>
<td></td>
<td>Orients objects</td>
</tr>
<tr>
<td>Pace items</td>
<td>Proceeds</td>
</tr>
<tr>
<td></td>
<td>Changes strategy</td>
</tr>
<tr>
<td></td>
<td>Flow in bimanual task performance</td>
</tr>
</tbody>
</table>

The AHA is only to be used by certified users. Certification requires a 2½-day course followed by a certification procedure involving independent scoring of eight assessments and filming three of one’s own cases, individual feedback on which is provided by course instructors.

Continued development of the AHA

Soon after the initial development the need to use the AHA for children over the age of five became evident. The idea of using an activity that was enjoyable for the children and interesting enough to elicit spontaneous hand use was crucial to the AHA. When letting older children play with the toys it was noted that some used their hand in a manner that was not the quickest and most efficient for them, but rather in a way that reflected a maximum use of the assisting hand. It was noted that some children thought they should use their affected hand in this way, since this was usually expected when visiting the OT. There was thus a need for an age-appropriate activity to elicit spontaneous hand use. We therefore developed two board games that were similar in concept, but had different themes. In both games the child travels on a trail, and at each stop on the way takes a mission card detailing a task that involves using one of the toys
in the AHA suitcase. The board game was developed in two versions. The first is based on an alien theme whereby the child has crash-landed on an alien planet and needs to find fuel for the spaceship to get back to earth (the Alien game, Figure 3a). The other version has an adventure theme whereby the child gets a mission to free a prisoner being held captive in a fortress (the Fortress game, Figure 3b).

These games were pilot-tested on 15 children aged six to twelve. It was found that the games had the intended effect; they were so enjoyable and captured the children’s interest so much that the children’s usual use of the affected hand was elicited. The spontaneous play on the part of the preschool children was called the Small Kids AHA, and the play using the board games was called the School Kids AHA. The items and scoring criteria described above were also used for the School Kids AHA.

A requirement for use of a test in clinical practice or research is that it have sound psychometric properties, such as validity and reliability, which should be reported by test developers (American Educational Research Association 1999; Law et al. 2005).

Validity

Validity is the most fundamental aspect of a test (American Educational Research Association 1999). It refers to whether a test measures what it intends to measure (Portney & Watkins 2000), or more specifically: "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9, American Educational Research Association 1999). In many textbooks different types of validity are presented (see DePoy & Gitlin 1999; Portney & Watkins 2000). Other authors, e.g. the American Educational Research Association in “Standards for educational and psychological testing” (1999), emphasize validity as a unitary concept for which there are different ways of gathering evidence. The aim of these ways is ultimately to provide evidence of the confidence that can be placed in the interpretations of results (American Educational Research Association 1999; Streiner & Norman 2003). An overview of the classic types of validity and their equivalence as defined by the “Standards” is presented in Table II.

A point crucial to validity is that a test should measure only one aspect of a phenomenon or construct and measure all aspects of that particular construct. Validity evidence based on test content refers to whether all items in the test represent the construct and whether all aspects of this construct are targeted in the items. This
evidence is usually gathered by consulting a panel of experts in the field or other people with experience of the construct being measured (American Educational Research Association 1999). The classical type of validity called face validity is also a way to gather evidence based on test content. Face validity means that on inspection test items seem to represent the construct one is intending to measure. This is a simple form of evidence and not enough to establish validity. However, it may be important both to the person being tested and to the test administrator that completion of the test is considered meaningful (Portney & Watkins 2000). Validity evidence based on internal structure is usually evaluated by analyzing whether items in the test form a unidimensional construct in a statistical analysis, for example in factor analysis, principal components analysis or item response theory, one example of which is Rasch measurement analysis (Streiner & Norman 2003). Validity evidence based on response processes is gathered by observing whether persons are scored on or answer the test as intended. This type of evidence also involves targeting of item difficulty to the ability of the persons tested (American Educational Research Association 1999). Validity evidence based on relations to other variables can be evaluated by comparing the results of the test in question with those of another test that claims to measure the same construct (concurrent evidence) or other criteria that are closely connected with the proposed construct. Predictive criterion evidence of validity refers to whether the test can be used to predict the occurrence of something (American Educational Research Association 1999).

**Rasch analysis to evaluate validity**

Rasch analysis is a method of developing and evaluating tests that has become widely used in the area of rehabilitation (Fisher 1993; Haley, Coster et al. 1992; Russell, Avery et al. 2000). It has its origin in a probabilistic model published by the Danish mathematician Georg Rasch in 1960 (Wright & Stone 1979). This model, along with developments thereof, together form a family of Rasch models. They are all based on the same basic theory, but are applicable to different types of data (Wright & Masters 1982; Wright & Mok 2000). The essence of the Rasch models is that the probability of success of an item depends only on the difficulty of the item and the ability of the person. In the model all items in a test are ordered from easier to harder. This ordering of items needs to be logical and follow the theoretical construct that it is intended to measure. Similarly, all persons can be ordered from less able to more able. Items and persons can be arranged along the same ability/difficulty line. This line represents the continuous measure line of the construct (Bond & Fox 2007). There are two underlying assumptions in the Rasch model:

- All persons are more likely to perform easy than difficult items.
- All items are more likely to be performed by persons of high ability rather than those of low ability (Bond & Fox 2007; Wright & Stone 1979).

When a person’s ability can be estimated to a point on this continuous measure line, a measure can be given to that person. A measure has the qualities of interval-level data. In the Rasch model a logarithmic transformation is used to transform raw scores into measures: log-odds information units (logits) (Bond & Fox 2007; Wright & Mok
Introduction

2000). The Rasch assumptions require that only one attribute or dimension be measured at a time: unidimensionality. However, this is only a theoretical idealization, since there are no completely unidimensional tests, but unidimensionality can rather be viewed as a continuum (Bond & Fox 2007; Linacre, 2006; Smith 2002).

Two models in the Rasch family are the partial credit model and the rating scale model. The partial credit model allows variation in the relative difficulty of the scale steps from item to item (Wright & Masters 1982). It requires a large sample of at least ten observations per rating scale category for a stable estimation (Linacre 1994). The rating scale model uses the same rating scale structure for all items and is preferable in smaller samples.

In Rasch analysis a number of concepts and ways of looking at data are provided. How these correspond to the validity evidence discussed above is presented in Table II.

Table II. Comparison of validity concepts between validity evidence, classical types, and validity concepts in Rasch analysis. (Adapted from Kottorp, 2003, p. 26, published with the author's permission)

<table>
<thead>
<tr>
<th>Validity evidence based on test content</th>
<th>Classical validity type</th>
<th>Rasch validity concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity evidence based on internal structure</td>
<td>Construct validity</td>
<td>Internal scale validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Principal Components Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rating Scale Structure</td>
</tr>
<tr>
<td>Validity evidence based on response processes</td>
<td>Not specified</td>
<td>Person response validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Targeting test item difficulties to ability of sample</td>
</tr>
<tr>
<td>Validity evidence based on relations to other variables</td>
<td>Criterion-related validity</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test linking</td>
</tr>
</tbody>
</table>

Internal scale validity evaluates of how well items in the scale fit the Rasch model assumptions and is expressed in terms of goodness-of-fit statistics. The fit statistics are calculated by subtracting the observed from the expected value for each observation, and they are summarized for each item in a mean square (MnSq) fit statistic and a standardized fit statistic (z). The fit statistics can either be calculated as the sum of the squared residuals, which produces the outfit statistics or as infit statistics, which is an information-weighted sum whereby unexpected responses from persons with ability closer to the item measure are given more emphasis. Item fit is a way of evaluating the extent to which each item meaningfully contributes to a unidimensional construct (Bond & Fox 2007). Another way of investigating unidimensionality is principal components analysis (PCA) of the standardized residuals. This is performed to examine whether a second dimension exists in the unexplained variance apart from the Rasch dimension (or the dimension defined by the construct). The comparison between the Rasch dimension and the first residual factor (largest single dimension in the unexplained variance) identifies possible multidimensionality (Bond & Fox 2007).
Introduction

The rating scale structure concerns the use of scale steps according to which each item is rated. Between the scale steps there are thresholds, e.g., a four-point rating scale has three thresholds. The threshold is defined as the measure level at which the probabilities of getting the higher and the lower score are equal. The threshold measures need to be ordered i.e., the threshold between Scores 1 and 2 has to have a lower measure than the next threshold between Scores 2 and 3 etc., otherwise the thresholds are disordered, indicating that the rating scale is not being used as intended. In a partial credit model, each item has its own threshold measures (Bond & Fox 2007; Linacre 2002; Tennant 2004; Wright & Stone 2004).

Person-response validity is an evaluation of how well the ability of each person fits with the assumptions of the Rasch model. The fit to the model is expressed as goodness-of-fit statistics which is comparable with item fit statistics. Another important factor regarding validity is whether the items in the test are well targeted to the ability of the persons for whom the test is intended (Bond & Fox 2007; Wright & Stone 1979).

Another crucial aspect of a scale is invariance. Invariance refers to that item difficulties should stay the same regardless of who takes the test (within the population that the test is intended for), and likewise that person measures stay the same regardless of which items are used (Bond & Fox 2007). One aspect of invariance pertaining to validity is differential item functioning (DIF). Items in a scale should function in the same way regardless of the attributes of the persons concerned (Bond & Fox 2007). Another application of invariance is test linking. Test linking can be an evaluation of both validity and reliability. Bond and Fox (2007) argue that common person linking is a way of evaluating whether two tests or two alternate forms of a test measure the same construct.

Reliability

Reliability expresses how stable and consistent an instrument is in measuring a trait, and the accuracy of that measurement (Polit & Hungler 1995). The reliability of a test has implications for validity. If measures produced are not stable the inferences drawn from the test scores may not be valid (American Educational Research Association 1999). Error is an unavoidable circumstance in all measurement and the crucial issue in reliability is estimating and reducing measurement error. In classical measurement theory, the role of error is explained with an equation where the observed score equals the true score plus error. Reliability is described as an estimation of how well the observed score represents the individual’s true score (Benson & Schell 1997). Error is considered random, and theoretically, if a large number of measurements are performed the error is considered to have a mean of 0 (Thorndike 1987). Usually a more serious problem than random error is systematic error, which needs to be minimized in any measurement situation. What is considered good reliability of a scale depends on how much the trait normally varies between individuals. An illustration of this would be the different requirements for a thermometer depending on whether it is to measure outdoor temperature or body temperature. For an outdoor thermometer a variability of \( \pm 1^\circ \text{Celsius} \) would normally be acceptable, which of course is not the case for a body thermometer (Streiner & Norman 2003).
There are a number of different aspects of reliability, one of which is *internal consistency*, which refers to the homogeneity of items, i.e. whether they correlate to each other and to the total test score (Streiner & Norman 2003). The estimate of internal consistency is based on the mean correlation between all items in the test (Nunnally & Bernstein 1994).

Another aspect of reliability is *stability*, which means to what extent a repeated measurement within a short period of time produces the same result as the first measurement. Day-to-day fluctuations in behavior affect all observational tests, and the evaluation of stability reveals the magnitude of expected fluctuations within the same person with maintained ability. Stability is tested using test-retest (retest), whereby the trial is repeated after a period of time short enough for the trait to be the same and long enough for the first testing not to influence the second one (Nunnally & Bernstein 1994; Portney & Watkins 2000).

The aspect of *equivalence* of a scale has two cases. It can be investigated by using two different versions of a test and comparing the results, which is referred to as parallel or *alternate forms* reliability (Polit & Hungler 1995). Equivalence also has the form rater reliability, which is applicable to tests involving an assessment of some sort. Rater reliability can either be between or within raters. Interrater involves different raters assessing the same situation with the same scale, and intrarater means the reliability of the same rater assessing the same situation repeatedly, with a time lapse sufficient to reduce the influence of memory (Streiner & Norman 2003).

*Statistical methods to evaluate reliability*

Reliability is often expressed as a reliability coefficient varying between 0 and 1 which “reflects the extent to which a measurement instrument can differentiate among individuals…” (p. 128, Streiner & Norman 2003) Analyzed further, the reliability coefficient is an expression of the proportion between the variance of the true score and the variance of the observed score. The reliability coefficient can be calculated in a number of ways depending on the aspect of reliability investigated.

The internal consistency is predominantly evaluated using Cronbach’s Alpha (or coefficient alpha) (Spector 1992). The value of the coefficient alpha is influenced both by the internal consistency and by the number of items, and it increases with the number of items in the scale (Nunnally & Bernstein 1994). In Rasch analysis the corresponding analysis is called person separation reliability, which is also expressed as a reliability coefficient (Bond & Fox 2007). The person separation reliability is related to the person separation index, which indicates to what extent different levels of ability can be distinguished using the test.

In calculation of the reliability coefficient for stability or equivalence, the first choice to be made is whether consistency or absolute agreement is of interest. Consistency refers to whether, for example, raters rank the individuals measured in the same ability order, whereas in absolute agreement the actual scores from each rater are compared (Streiner & Norman 2003). When the reliability coefficient for consistency is calculated using
Introduction

the Pearson product moment correlation, it should be interpreted with caution, since it reflects to what extent assessments from two raters form a straight regression line, and the coefficient could be 1 even though there are considerable differences in actual scores. In this analysis only two e.g. raters can be compared at the time (Streiner & Norman 2003).

To calculate the reliability coefficient for agreement, the level of data produced by assessments influences what kind of analysis to choose. A review of 57 articles evaluating stability or equivalence in assessments of motor skills showed that the Intraclass Correlation Coefficient (ICC) was the most commonly used analysis (unpublished review by this author). The ICC is recommended for continuous data and the Kappa analysis for dichotomous ordinal level data (Switzer, Wisniewski et al. 1999). To analyze a test with more than two scale steps the Kappa can be weighted, thus giving credit to a partial agreement. The most common weighting scheme is quadratic weights (Streiner & Norman 2003). The ICC produces the same coefficient as the weighted Kappa with quadratic weights. The calculation chosen could depend on practical considerations (Fleiss & Cohen 1973).

The ICC was comprehensively described by Shrout and Fleiss in an article from 1979, and they present different forms of the ICC depending on the conditions. There are three main conditions: 1) Each individual is rated by a group of raters randomly selected from a larger population of raters; 2) A group of individuals is rated by a group of raters randomly selected; 3) Each individual is rated by all raters, who are the only raters in question. Each condition corresponds to a model of analysis of variance (ANOVA): 1) One-way random effects model; 2) Two-way random effects model; 3) Two-way mixed effects model (Shrout & Fleiss 1979).

What constitutes an adequate enough reliability coefficient has been debated. In general, reliability requirements are more stringent for decisions about an individual than for evaluations at group level, since the calculation of average scores reduces measurement error (Nunnally & Bernstein 1994). A rule of thumb is that for group comparisons a coefficient of 0.70, or even better 0.80, is desirable. When making decisions about individuals a coefficient of at least 0.90, and preferably 0.95, is desirable (Nunnally & Bernstein 1994). The Kappa produces a lower coefficient than the ICC, and accordingly the classic standards reported by Landis and Koch (1977) are lower than the above.

A difficulty with the reliability coefficient is that it does not reveal the error connected with an individual’s test score. The Standard Error of Measurement (SEM) is a complementary way of reporting reliability which expresses measurement error in the same unit as the measure. The SEM can be used to calculate a 95% confidence interval (CI) of ±2SEM around a single sum score (Thorndike 1987). Another measure of the error around a score is the standard error (SE) which is calculated for each person and item measure in a Rasch analysis (Bond & Fox 2007).

The calculation of SEM opens up the possibility of calculating the minimum measurable change between two test occasions. The intuitive way would be to use the
two CIs above, though Eliasziw and colleagues argue that this calculation might overestimate the measurement error (Eliasziw, Young et al. 1994). Eliasziw provides a way of calculating whether a change between two measurements is statistically significant using the SEM. The same reasoning can be used to calculate the Smallest Detectable Difference (SDD), which is also derived from the SEM. The SDD is the smallest change that must take place between two measurements for the test to detect statistically real change (Beckerman, Roebroeck et al. 2001; Schreuders, Roebroeck et al. 2003). SDD has also been referred to as minimal detectable change (MacDermid & Stratford 2004) or smallest real change (Beckerman et al. 2001).

Initial evidence of psychometric properties of the AHA

In the paper describing the development of the AHA (Krumlinde-Sundholm & Eliasson 2003), 60 assessments of children with unilateral CP or OBPP were analyzed using Rasch analysis. Results showed evidence of internal scale validity. 95% of items fit the Rasch assumptions, thus supporting unidimensionality. Person response validity showed that 97% of the person’s responses fitted the model. Excellent targeting of item difficulty to the person’s ability was illustrated in a figure. A person separation index of 6.16 showed a very good ability to distinguish between children of different ability levels. These initial results showed that the AHA has the potential of becoming a useful test, because of its unique purpose and the good preliminary psychometric properties. Further development was suggested, e.g. evaluating the psychometric properties with a larger sample, inter- and intrarater reliability and extending the use of the AHA to older children (Krumlinde-Sundholm & Eliasson 2003).

CHILDREN WITH A UNILATERAL DISABILITY

Children with two types of unilateral impairments were included in the studies in this thesis: unilateral cerebral palsy (CP) and obstetric brachial plexus palsy (OBPP).

Cerebral palsy (CP)

Cerebral palsy (CP) is the most common cause of physical disability in children. A number of different definitions of CP have been proposed over the years. The most recent one, developed in consensus between a number of authorities in the field, is:

“...a group of permanent disorders of the development of movement and posture, causing activity limitation, that are attributed to non-progressive disturbances that occurred in the developing fetal or infant brain. The motor disorders of cerebral palsy are often accompanied by disturbances of sensation, perception, cognition, communication, and behavior, by epilepsy, and by secondary musculoskeletal problems.” (p. 9, Rosenbaum, Paneth et al. 2007)

Owing to the variation in severity and type of involvement among children with CP there is also a need to further classify children within the group. Recommendations for classification are made in the above-mentioned document (Rosenbaum et al. 2007). In this classification the anatomic distribution is given i.e. unilateral or bilateral motor involvement (implying a departure from the classical terms hemiplegia, diplegia etc.). The dominant type of muscle tone is also described. The most common type of tone is
Introduction

spasticity (90%), which is characterized by increased reflex contraction when the muscle is stretched. About 6% of the children have dyskinetic symptoms, which means fluctuating muscle tone, and another 4% have ataxic symptoms, which are characterized by tremor (Brown & Walsh 2000; Krägeloh-Mann & Staudt 2008). Furthermore, the functional consequences of the CP in daily life need to be classified. The Gross Motor Function Classification System (GMFCS) is used to classify walking and mobility in five levels. A child in level I can walk independently on most surfaces and a child in level V is totally dependent, even when using an electric wheelchair (Palisano, Rosenbaum et al. 1997). The Manual Ability Classification System (MACS) – the fine-motor equivalent to GMFCS – is used to classify ability to handle objects in daily life. The MACS also has five levels, ranging from level I, where the child handles most objects in daily life easily and successfully, to level V, where the child needs constant assistance and can only operate adapted equipment such as large switches with effort (Eliasson, Krumlinde-Sundholm et al. 2006). Thereafter, the recommendation is to describe other impairments such as epilepsy and visual impairment (Rosenbaum et al. 2007). In this thesis this terminology is used, but the affected hand of children with unilateral CP will be sometimes referred to as the hemiplegic hand.

The overall prevalence of CP in Sweden is about 2-2.5 per 1,000 live births (Hagberg, Hagberg et al. 2001; Himmelmann, Hagberg et al. 2005; Westbom, Hagglund et al. 2007), and is similar across Europe (2.08/1000) (Surveillance of Cerebral Palsy in Europe (SCPE) 2002). The prevalence for babies born at full term is around 1-1.3/1,000 live births (Hagberg et al. 2001; Himmelmann et al. 2005; Himpens, Van den Broeck et al. 2008). For children born prematurely the prevalence is much higher: 146/1,000 for children born before 27 weeks of gestation with a significant decrease thereafter (Himpens et al. 2008; Surveillance of Cerebral Palsy in Europe (SCPE) 2002).

The brain abnormalities leading to CP are different in origin, and the type of lesion depends mainly on the state of maturation of the brain at the time of insult (Staudt, Gerloff et al. 2004). There are three major origins of CP. The most common cause (56%) is periventricular lesion (PVL), which occurs early in the third trimester (Krägeloh-Mann & Horber 2007). During this period the white matter surrounding the ventricles located in the centre of the brain is particularly sensitive to lack of oxygen and bleeding. Insults may result in reduced white matter around the ventricles or damaged and scarred white matter in the periventricular area. A PVL can occur just after birth in babies born very prematurely, or before birth in babies born at full term (Johnston & Hoon 2006; Krägeloh-Mann 2004; Krägeloh-Mann & Horber 2007). The second common origin of CP (18%) is cortical and/or deep grey-matter lesions late in the third trimester. In this period the grey-matter areas are vulnerable to lack of blood supply. Cortical/subcortical lesions can be caused by bleeding or infarct, as in adults. They may also be the result of extended lack of oxygen during birth, but this is far less common than previously thought. (Bax, Tydeman et al. 2006; Krägeloh-Mann 2004; Krägeloh-Mann & Horber 2007). A third cause of CP is maldevelopment (9%) which occurs during early brain development in the first 20 weeks of gestation. An insult or maldevelopment during this period leads to malformations in the brain structure, e.g.
Introduction

excessive or insufficient gyration or altered nerve cell layers in the cortex (Krägeloh-Mann 2004). The origins of maldevelopment are largely unknown: possible origins that have been discussed are genetic factors, infection and disturbance in the blood supply (Krägeloh-Mann 2004). The correlation between location of lesion and distribution of CP has been known for some time (Wiklund & Uvebrant 1991). More recently a relationship between size of lesion and severity of impairment has been established (Hashimoto, Hasegawa et al. 2001; Staudt, Niemann et al. 2000). In children with CP, motor impairments are more marked than sensory ones (Arnould, Penta et al. 2007).

Unilateral Cerebral Palsy

About a third of the children with CP have unilateral motor disabilities. Reports vary between 29% and 38% of the total CP population (Hagberg et al. 2001; Himmelmann et al. 2005; Surveillance of Cerebral Palsy in Europe (SCPE) 2002; Westbom et al. 2007). Most of the children with unilateral CP are born at full term, and about a third are born prematurely (Himmelmann et al. 2005). In 75% the origin is prenatal, the rest being perinatal (during birth until seven days after birth) (usually in premature births). Most children with unilateral CP have a brain lesion only in the side opposite to the paresis, some children have a bilateral lesion (12-24%), and in some children (about 10%) with mild symptoms no lesion can be seen in magnetic resonance imaging (Krägeloh-Mann & Horber 2007).

The way the hands are used is affected by many factors in the environment and the task and within the person. Within the person there are different factors that together influence the use of the hands: the cognitive ability, visual function, motivational factors etc. The degree of impairment varies from child to child, owing to differences in type and extent of brain lesion, and this has been demonstrated both in general for the whole CP group (Melhem, Hoon et al. 2000; van Haastert, de Vries et al. 2008) and for upper limb function in children with unilateral CP in particular (Cioni, Sales et al. 1999; Staudt et al. 2000). The specific impairments such as impaired motor and sensory functions in the hand are reported below.

The predominant muscle tone in unilateral CP is spastic. This increased muscle tone leads to more deliberate, slower movements and to weakness (Brown & Walsh 2000; Uvebrant 2000). These two factors, which are often mentioned as being the most disabling are also a sensitive indicator of the severity of hand disability (Brown & Walsh 2000). The ability to feel pain, temperature, position and vibration is often maintained (Brown & Walsh 2000), whereas stereognosis, 2-point discrimination and graphesthesia are impaired in a large group (Krumlinde-Sundholm & Eliasson 2002; Majnemer, Bourbonnais et al. 2008). Impaired sensation has been shown to correlate to poor dexterity (Krumlinde-Sundholm & Eliasson 2002), and leads to decreased ability to use sensation to determine the power needed for grasping. Typically, children with decreased sensation use more force than is necessary, to make sure they do not drop the object being grasped (Eliasson, Gordon et al. 1995). This may be a problem when handling fragile objects. Children with unilateral CP need more time to grasp and lift an object than typically developing children. Control of force in the affected hand in children with unilateral CP does not develop to a grip/lift synergy, but they keep the
pattern of forces increasing in sequence. This leads to a slower and more deliberate movement pattern (Eliasson, Forssberg et al. 2006; Eliasson, Gordon et al. 1991). The same prolonged time and sequencing occurs when putting down and releasing an object. The lack of grip/lift synergy correlates to poor manual dexterity (Forssberg, Eliasson et al. 1999).

Mirror movements are quite common in children with unilateral CP (Kuhtz-Buschbeck, Krumlinde-Sundholm et al. 2000). Minor mirror movements or none at all are found in children who have either a good or a very poor hand-function outcome (Carr, Harrison et al. 1993). Marked mirror movements are an indication that the undamaged hemisphere is controlling both hands (Staudt et al. 2004). Some control of the affected hand is of course an asset, and symmetrical bimanual movements can enhance the performance of the affected hand (Volman, Wijnroks et al. 2002). However, movements in the affected hand slow down and inhibit movements in the dominant hand in performance of asymmetric bimanual tasks, thus mirror movements can be problematic (Volman et al. 2002).

These factors jointly contribute to decreased function in the arm and hand, affecting children’s ability to perform everyday tasks. The degree to which hand function is affected varies greatly from child to child, and children with unilateral CP are predominantly found in MACS levels I-III.

Obstetric Brachial Plexus Palsy (OBPP)

A minority of the participants in the different studies in this thesis had an OBPP. An obstetric brachial plexus injury is a result of excessive traction of the spinal nerves of the brachial plexus during birth, first described by Duchenne (1872) and Erb (1874). This occurs in about 2/1000 live births and is followed by a residual palsy in 25% (Eng, Koch et al. 1978; Michelow, Clarke et al. 1994; Narakas 1986). The brachial plexus consists of five spinal nerves (C5-Th1) which contain all the sensory and motor nerves to the hand and arm. The extent of the injury depends on the number of nerves that have been damaged and the severity of the damage. An upper injury (Erb’s palsy) involving the spinal nerves C5-C6 (possibly also C7) and thus affecting shoulder movement and elbow flexion (possibly also extension) is most common (Gilbert 2001). When all five spinal nerves are involved the hand is also affected. The number of nerves involved can be observed just after birth. The severity of the damage to the nerves can only be determined by monitoring the rate of spontaneous recovery over time. In mild injuries the nerves may only be affected by an oedema, which will heal spontaneously in a few weeks. In the more serious cases with nerves disrupted or avulsed, there is no spontaneous healing. In cases with one or more spinal nerves torn, microsurgical nerve reconstruction during the first year of life can bring about some healing (Gilbert 1995; Narakas 1986). There is evidence of significant increase in shoulder movement following surgical reconstruction in Erb’s palsy (Strömbeck, Krumlinde-Sundholm et al. 2000). Secondary surgery with muscle and tendon transfers or osteotomies later on in childhood is common.
The most common symptoms in children with Erb’s palsy is limited in shoulder movement, in particular limited external rotation of the shoulder and extension of the elbow. The typical presentation is the “trumpet sign”, which leaves the elbow raised and shoulder extended when the hand is lifted to the mouth (Krumlinde-Sundholm, 2002). Sensory function in the hand is generally good in children with Erb’s palsy (Krumlinde-Sundholm, Eliasson et al. 1998). In a total plexus palsy all the movements in the arm and the hand are severely impaired, with minimal ability to move the wrist and fingers or none at all (Krumlinde-Sundholm, 2002).

The ability of children with OBPP to manage everyday tasks is usually good, but these children also display individual variation. They commonly become independent in their everyday activities, and usually learn to find alternative strategies so as to compensate for limited in hand function (Ho, Curtis et al. 2006).

DEVELOPMENT OF HAND FUNCTION

Typically developing children

In typically developing children hand skills develop rapidly an early age, with initial development of grasps occurring during the very first year of life (Gesell, Halverson et al. 1940). Hand movements in newborn babies are characterized by general movements and the reflex grasp (Twitchell 1970). Recent research has also shown the presence of goal-oriented movements in newborn babies (Nagy, Compagne et al. 2005; van der Meer, van der Weel et al. 1995). At about four to five months a voluntary whole-hand grasp develops, entailing increasing involvement of the thumb followed by the precision and pincer grasps at nine to ten months. After the first year, grasps are further refined and new patterns emerge (see Exner 2005). One example of this refinement is the development of force control in grasping. It starts with a sequential pattern (i.e. increased grip and load force, one after the other), which is mainly dependent on sensory and proprioceptive feedback. The coordination of forces develops during the preschool period, achieving a robust synergy at six to eight years of age (Forssberg, Eliasson et al. 1991). Young children also use a feedback strategy to regulate the amount of force needed to lift an object, and this is a slow strategy. There is a slow shift from a feedback to a feed-forward strategy in young children, resulting in quicker and more efficient control of force and speed. Adult-like patterns are reached at the age of about eight to ten (Forssberg, Kinoshita et al. 1992).

Typically developing children use both their hands together symmetrically from early age, and at about three to four months they hold an object with both hands together (Fagard & Jacquet 1989). During the second half of the first year (about eight months) the child starts to hold an object with one hand while the other is exploring it (Newell & McDonald 1997). Complementary movements involving different hand roles starts to develop during the second year of life (Fagard & Jacquet 1989). Young children find it easier to perform mutually mirroring bimanual movements than identical movements. The ability to perform identical or asymmetrical bimanual movements continues to develop until the age of about ten (Fagard, Hardy-Léger et al. 2001; Gordon & Steenbergen 2008).
Introduction

Development in children with CP

As discussed above, hand function in children with CP is impaired in different ways, but there are few studies of the course of development of hand function in these children. Thus all the studies found that concern development in children with all types of CP are discussed below not just those concerning children with unilateral CP. To my knowledge the only growth curves regarding hand function were presented by Hanna and colleagues in 2003. They reanalyzed data from an earlier study in which 51 children with CP aged 16 months to five years were each measured four times over a ten-month period. They used the Peabody fine-motor scales (Folio & Fewell 1983), which measures grasp and visual-motor integration, and the Quality of Upper Extremity Skills Test (DeMatteo, Law et al. 1992), which measures quality of movement. The two tests showed somewhat different development, with a steady increase in average ability in Peabody, but with increase, peak and decline in QUEST. This difference is attributed to the fact that QUEST measures quality of movement, and children's range and control of movement may decrease as they develop, owing to spasticity and contractures. Peabody, on the other hand measures abilities that are likely to improve with practice, e.g. handling a pen and paper. Results showed that different distributions (hemiplegia or quadriplegia) and degrees of CP severity (mild, moderate or severe) could predict outcome.

Another recent study which covered a long period of time was conducted by Eliasson et al. (2006a). They reported results from development in higher ages from a follow-up of ten people with CP, who were first measured at ages 6-8 years, and then at 19-21 years. The participants had large variation of impaired hand function. They used the Jebsen-Taylor test of hand function, which is a set of timed tasks, and an experimental design using a grip-lift task whilst measuring force control. They found that the group as a whole improved their time for completion of the tasks in the Jebsen-Taylor test by 45%. On the grip-lift task the time taken to prepare and lift an object decreased by 22%, demonstrating consistent and efficient improvement.

Smits-Engelsman et al. (2004) studied grip force in a cross-sectional experimental study of 20 children with mild unilateral CP and an age-matched control group. As expected, in the study group grip force was weaker in the affected hand than in the non-affected hand. In both hands the grip force was smaller for the older children (10-15 yrs) than for the younger group (5-10 yrs) (Smits-Engelsman, Rameckers et al. 2004). This would imply that grip force in both hands deteriorates during the school years in children with unilateral CP. This has to my knowledge not been described elsewhere. Owing to the cross-sectional design and the small study group, these differences may be due to the study design only and therefore need to be confirmed by a larger longitudinal study in which the same children are followed over time.

Development of hand function was also studied by Fedrizzi et al. (2003) in 31 children with unilateral CP. They measured quality of grip and spontaneous hand use in handling objects from the age of four until after the age of 11. Results showed a weak trend regarding improvement in grip patterns during the period, although it was not significant. They found no changes in spontaneous hand use during the period. The
measures used were single-item, non-standardized and unvalidated, thus no confidence can be placed in the implication of these results.

In conclusion, the above studies suggest that children with mild CP display positive development of both force control and functional use of the hand, and in children with more severe CP some increase in capacity for movement followed by deterioration in capacity may be seen.
AIMS
The general aim of this thesis was to evaluate the psychometric properties of the AHA for an expanded age range and to use the AHA to describe development of assisting hand use in children with unilateral CP.

SPECIFIC AIMS

• To evaluate the validity of the AHA for the expanded age range 18 months to 12 years. (Study I and supplementary analysis of data in Study I for this summary)

• To evaluate the reliability of the AHA with regard to:
  • Interrater and intrarater reliability (Study II)
  • Retest reliability of the Small Kids and School Kids AHA (Study III)
  • Alternate forms reliability between the Small Kids and the School Kids AHA and between the Alien game and the Fortress game in the School Kids AHA (Study III)

• To describe longitudinal development of the usefulness of the assisting hand in children with unilateral CP between the ages of 18 months and eight years (Study IV).
MATERIAL AND METHODS

DESIGN

In Studies I-III the psychometric properties of the AHA were evaluated. In Studies I and III the AHA was extended to the age range 18 months - 12 years. In Study IV the AHA was used to describe development of the usefulness of the hemiplegic hand in children with unilateral CP (Table III). Data in all studies consisted of assessments with the AHA.

Table III. Aims, designs and methods for data analysis for the studies in the thesis.

<table>
<thead>
<tr>
<th>Study no</th>
<th>Aim</th>
<th>Design</th>
<th>Data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Validity for extended age range</td>
<td>Cross-sectional Instrument evaluation</td>
<td>Rasch analysis</td>
</tr>
<tr>
<td>II</td>
<td>Rater reliability</td>
<td>Cross-sectional Instrument evaluation</td>
<td>ICC, SEM Spearman’s rank order correlation</td>
</tr>
<tr>
<td>III</td>
<td>Retest and alternate forms reliability</td>
<td>Cross-sectional Instrument evaluation</td>
<td>ICC, SEM, SDD t-test, Wilcoxon matched pairs signed ranks test Bland-Altman plot Test linking in Rasch analysis</td>
</tr>
<tr>
<td>IV</td>
<td>Longitudinal development of assisting hand use</td>
<td>Prospective longitudinal</td>
<td>Non-linear mixed models</td>
</tr>
</tbody>
</table>

PARTICIPANTS

This thesis is based on AHA assessments of 367 children with unilateral CP or OBPP (Table IV).

Table IV. The distribution of participants and number of AHA sessions.

<table>
<thead>
<tr>
<th>Study no</th>
<th>Aim</th>
<th>Children (n)</th>
<th>Diagnosis CP (n)</th>
<th>OBPP (n)</th>
<th>AHA sessions (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Validity and reliability for age range 18 months – 12 yrs</td>
<td>303</td>
<td>256</td>
<td>47</td>
<td>409 (2-5 for 48 children)</td>
</tr>
<tr>
<td>II</td>
<td>Interrater reliability 2-rater design</td>
<td>18</td>
<td>16</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>III</td>
<td>Retest reliability Small Kids vs School Kids</td>
<td>18</td>
<td>18</td>
<td>0</td>
<td>36 (2 per child)</td>
</tr>
<tr>
<td>IV</td>
<td>Longitudinal development of assisting hand use</td>
<td>43</td>
<td>43</td>
<td>0</td>
<td>280 (mean 6.5 per child)</td>
</tr>
</tbody>
</table>
Material and Methods

Some children participated in more than one study, and the overlap is reported in Table V. Children for Studies I, II and IV (n=319) were recruited via OTs at local habilitation centres or hospitals in Sweden. For Study III, 48 new children with unilateral CP were recruited via local (re-)habilitation centres or special schools in Australia, the Netherlands and Sweden.

Table V. Number of participants overlapping between studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Study I (n=303)</th>
<th>Study II a (n=18)</th>
<th>Study II b (n=8)</th>
<th>Study III (n=55)</th>
<th>Study IV (n=43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study II a</td>
<td>13</td>
<td>6</td>
<td>2 a</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>Study II b</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Study III</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Study IV</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Study II a=the interrater design with 2 raters in Study II.
Study II b=the inter- and intrarater designs with 20 raters in Study II.

a Different AHA sessions with these two children were used for the two parts of Study II.

MEASURES AND DATA COLLECTION

In all studies the AHA was used to collect data. In Study IV the MACS was used to classify the children.

The AHA-films used in Studies I and II were previously recorded and collected for other research projects (Eliasson, Krumlinde-Sundholm et al. 2005; Krumlinde-Sundholm & Eliasson 2003, Study IV in this thesis) and in clinic. Some children were filmed more than once, up to 5 times (Table IV). In Study I all eligible assessments available in the collection at the time were included (n=409 in ages 18 months – 12 years). In Study I all assessments were performed by one of two experienced raters. For this study no new assessments of the films were conducted, but existing scores were used in the analysis.

In Study II, two different sampling strategies were utilized to select films of children aged 18 months to 5 years from the collection of AHA-films mentioned above. In the 2-rater interrater design a standard randomisation was used to select 18 films and two raters independently assessed these AHA-sessions. For the 20-rater design, the films available were stratified according to ability level and equal numbers of AHA sessions were randomly selected from each stratum. For the 20-rater inter- and intrarater designs, 10 Swedish-speaking and 10 English-speaking raters were recruited. Raters were asked to participate in consecutive order of certification and the first 10 from each language group were included. All 20 raters assessed the same eight AHA sessions, and each rater made a repeated assessment of one play session a minimum of three weeks after the first assessment.

In Study III three groups of children were recruited (Table IV). In the first part children aged 18 months to 5 years participated in the retest evaluation of the Small Kids AHA. The children in the second group (5 – 6 years old) were recruited to evaluate alternate forms reliability between Small Kids and School Kids AHA. The third group, aged 6 –
12 years, participated in the trials of both retest reliability of the School Kids AHA and alternate forms reliability between the two games in School Kids AHA.

In Study III all children were filmed twice, at an interval of approximately two weeks interval. In the first part of the study – retest of the Small Kids AHA – the children were tested repeatedly using the Small Kids AHA. In the second part – alternate forms reliability of the Small Kids vs. School Kids AHA – each child was tested using both the Small Kids AHA and the School Kids AHA. Nine children did the School Kids AHA in the first session, of which four played the Alien game and five the Fortress game. The other nine children did the School Kids AHA in their second session, five of them playing the Alien game and four the Fortress game.

In the third part of Study III, both retest and alternate forms reliability of the School Kids AHA were evaluated. For the alternate forms evaluation each child was tested using both board games with about two weeks apart. Eleven children played the Fortress game in the first session and eight children started with the Alien game. To evaluate retest reliability the plan was to do a third session two weeks after the last one with the same game as in the last session. This was not, however, considered necessary if the two first assessments produced similar results. Therefore after completion of the first two sessions the therapist collecting the data made preliminary assessments and if the raw score difference exceeded 3 points a third session would be conducted. The difference was ≤3 points for all children in this group. Thus the retest reliability was calculated from the alternate forms data collection. All assessments for the analysis in Study III were performed by one rater.

In Study IV the protocol was to measure each child for a minimum of three years. Children under the age of 3 were measured every six months, and subsequently once a year. Owing to practical circumstances there were a number of exceptions to this protocol. Children were included at different ages with a mean inclusion age of 2 years 8 months. AHA sessions were conducted by the researchers or by local therapists. At the time of the last AHA session, information about the child’s MACS level, if it had been determined, was collected from the child’s occupational therapist. If the MACS level was not available it was determined by interviewing the child’s parents. To reduce rater error, all AHA sessions with one child were assessed by the same rater. In all, three raters were involved in the assessment for Study IV. To further reduce intrarater error, all play sessions for each child were assessed on the same day or on consecutive days. The sessions of one child were watched in random order and once all sessions were rated the scores were inspected. Where there was a variation in scores the films were re-observed to find out whether the variation was due to an actual change in the child’s behaviour or to rater error.

STATISTICAL ANALYSIS

A range of statistical methods was used in this thesis (Table III). The three main methods (Rasch analysis, ANOVA-based analysis and mixed models) are described below. In addition to these three, a few other methods were used. In Study II the Spearman’s rank order correlation was calculated to evaluate whether experience of the
raters was correlated to their scoring accuracy. The rater’s experience was defined both as the number of assessments before the study and as the time gap between the instructional course and the study. Their accuracy was defined as the deviation both from the group median and from a golden standard. In Study III a paired t-test was used to evaluate whether there was a systematic difference between test and retest and between the different versions of the AHA. The Wilcoxon matched pairs signed ranks test was used to evaluate the differences between sessions and versions at item level (Altman 1991). To illustrate the difference in sum score between sessions and versions Bland-Altman plots were created for raw score sums. In the Bland-Altman plot the difference between test times (or versions) was plotted against the mean sum score for each person. The limits of agreement were calculated as the mean difference for the group ±2SD (Bland & Altman 1986). The SPSS for Windows software was used for the above and for ANOVA-based analysis.

Rasch-based analysis

Rasch analysis was used to evaluate the validity and one aspect of reliability of the AHA in Study I, to convert raw scores to logits in Studies III and IV and to conduct test linking in Study III. During the period since analysis for Study I there has been some development in the area of Rasch analysis along with upgraded and more accessible software enabling more detailed and informative analysis. Thus the analysis of data in Study I was supplemented by a new analysis for this summary, using the same data set. The analysis in Study I was conducted in the Facets 3.11 software and the new analysis as well as analyzes for Studies III and IV in Winsteps 3.65. The Rasch-based analyzes were mainly performed using a partial credit model, with the few exceptions stated below, where the rating scale model was used.

Internal scale aspects of validity (Study I)

In Study I evidence of internal scale aspects of validity was evaluated by inspecting infit MnSq and z-values. The MnSq can range from 0 to ∞ and has an expected value of 1. The z-statistic has an expected value of 0. For Study I in this thesis Infit MnSq ≤1.4 in combination with z-values <2 were set as criteria for acceptable item fit (Wright & Linacre 1994). The infit statistic was chosen since it is more sensitive to misfitting items close to the person’s ability, but both infit and outfit statistics were reported. Smith (1991) found that even with a perfect fit to the model, 5% of items will display misfit by random. A 95% fit was thus considered as evidence to support item fit to the model and unidimensionality.

Principal components analysis (supplementary analysis for this summary)

Data from Study I was analyzed in a principal components analysis (PCA) of the standardised residuals to further evaluate to what extent the items measured a single dimension. Criteria for unidimensionality were that at least 60% of the raw variance should be explained by the dimension of the measure and not more than 5% of the variance should be in the largest secondary dimension (Linacre, 2006).
Material and Methods

Rating scale structure (supplementary analysis for this summary)

The further analysis for this summary included another aspect of validity within the scale: ordering of scale categories. This was done to determine whether the rating scale for each item was being used in an expected manner. The threshold measures for the 22 items in the AHA were calculated and inspected.

Person response aspects of validity (Study I)

In Study I, infit MnSq and z-statistics were used to evaluate person fit from the same criteria as used for item fit. A 95% fit was considered as evidence to support person response aspects of validity. Furthermore, targeting of test items to the ability of the persons measured was evaluated. In Study I this was explored by comparing the mean item measure to the mean person measure. In this summary the targeting was illustrated with a graph.

Differential item functioning (supplementary analysis for this summary)

Another analysis that was conducted on data in Study I for this summary was a DIF analysis. This was to determine whether some items are easier or more difficult to perform for subgroups in the sample. Subgroups evaluated were diagnosis and age group. There was an uneven distribution in the sample between assessments of children with unilateral CP (n=362) and OBPP (n=47). To level out this difference a subsample of 46 assessments from the CP-group was randomly selected. Since this smaller sample produced a poor representation in some scale categories a rating scale analysis was conducted. For the age group comparison the sample of assessments from younger children (age ≤60 months) was reduced by random selection to 147 assessments to match the 147 assessments from children aged >60 months. Item measures for each subgroup were calculated and compared in a plot where the 95% confidence interval lines were also drawn. Items outside these intervals were considered as DIF (Bond & Fox 2007). To further evaluate whether DIF-items produced biased item measures the person measures from the full 22-item scale were compared with person measures from the scale with DIF items deleted (Roznowski & Reith 1999; Woodbury, Velozo et al. 2008).

Test linking (Study III)

Common item and person linking was performed between alternate versions of AHA in Study III. This was done to evaluate whether scores from different test versions are equal and whether different versions measure the same construct. Separate Rasch analyses for each version of the AHA were run to obtain item and person measures. Owing to the small sample in this study the rating scale model was used. Item and person measures for Small Kids vs School Kids AHA and for Alien vs Fortress board games were plotted against each other in separate plots. For test equivalence 95% of items should fall within the 95% confidence bands that are drawn in the plots (Bond & Fox 2007). In the common person linking plot relative difficulties of the tests are compared by inspecting the empirical line, which should go through the origin (Bond & Fox 2007).
**Material and Methods**

*Person separation: reliability and index (Study I)*

In Study I person separation reliability and index were calculated. The person separation reliability is expressed as a coefficient that can vary between 0 and 1. The separation reliability is related to the separation index (G), from which number of statistically different ability strata was calculated using the formula (Fisher Jr 1992; Wright 1996):

\[
\text{No of strata} = \frac{(4G+1)}{3}
\]  
(1)

In order for the scale to meaningfully separate individuals it must separate two distinctly different ability strata. This is equivalent to a person separation reliability of 0.7 (Fisher Jr 1992).

**ANOVA-based analyses**

*Reliability coefficients (Studies II and III)*

Reliability coefficients were calculated to evaluate to what extent the test can distinguish between individuals of different ability. As a reliability coefficient, the ICC was calculated in the evaluation of inter- and intrarater, retest and alternate forms reliability (Studies II and III). For inter-rater reliability ICC 2,1, or a two-way random effects model, was used and for intra-rater, retest and alternate forms reliability a one-way random effects model, or ICC 1,1 was used (Shrout & Fleiss 1979).

*Standard Error of Measurement (Studies II and III)*

To obtain an error range within which the true score would lie with 95% certainty the SEM was calculated for rater reliability. The SEM was calculated for the interrater condition, and data from the 2-rater design was used. SEM was also calculated for the intrarater condition and here assessments of eight raters – one for each of the eight children in the trial – were used. The SEM was also calculated for the different designs in Study III. A one-way ANOVA-analysis with time as fixed factor and child as random factor was used to obtain the SEM with the formula:

\[
\text{SEM} = \sqrt{(SS/df)} = \sqrt{(SS_{time} + SS_{child*time})/ (df_{time} + df_{child*time})}.
\]  
(2)

The error interval was calculated as ±2 SEM (Thorndike 1987).

*Smallest Detectable Difference (Study III)*

The SDD was calculated from the SEM in the cases where repeated testing was involved: retest and alternate forms reliability. The following formula was used:

\[
\text{SDD} = 1.96 \times \sqrt{2} \times \text{SEM}
\]  
(3)

The SDD is interpreted as the smallest difference between two test sessions that the test can distinguish between. A smaller difference than the SDD is not considered a measurable change, but is within the error range inherent in the test (Beckerman et al.}
The significance level of the score change for the SDD was calculated using the formula from (Eliasziw et al. 1994):

$$Z = \frac{\text{measure}_1 - \text{measure}_2}{\sqrt{2 \text{ SEM}}}$$

The p-value corresponding to the test statistic (Z) was found in a standard normal distribution table.

Mixed models for longitudinal data (Study IV)

Longitudinal data analysis requires repeated measures of the same persons, ideally $\geq 3$ measures per person (Singer & Willet 2003; Weiss 2005). The goal of a longitudinal analysis is to obtain a mathematical function which explains as much as possible of the variation in data using as few parameters as possible. This function (if based on a representative sample) can then be used to predict outcome for other individuals (Singer & Willet 2003). In Study IV a non-linear mixed model was used to analyze longitudinal data. The choice of model was based on theoretical considerations of how fine motor function develops both in general and in children with unilateral CP in particular, as well as on empirical data. A negative exponential shape of the non-linear model was chosen with three parameters: start, maximum level (limit) of AHA and rate (how rapidly the limit is reached). The higher the value of the rate parameter, the faster the limit is reached (Singer & Willet 2003). The function used was:

$$\text{AHA} = \text{Limit} - (\text{limit} - \text{start}) \cdot e^{\text{rate} \cdot \text{age}}$$

In this study the start level was assumed to be -10.18 logits, which is the minimum AHA logit score, at age 0 months for all children. The limit and rate were set as both random and fixed effects in the model, which reflects that all children were allowed their own individual limit and rate. Individual levels of limit and rate were used to calculate the average limit and rate for the group. There was considerable inter-individual variability both in limits and rates among the 43 children in the group. It was thus considered meaningful to divide them into subgroups. Two models for grouping of children were chosen: AHA score at 18 months and MACS level. The AHA scores at 18 months (observed, and if not available; predicted) displayed a natural divide in the empirical data around -3 logits. When looking more closely at the performance on AHA items for these children it became clear that the children with a score level below -3 logits did not use a grip with the assisting hand, whereas the children with an AHA score above -3 logits used a grip to hold objects with the assisting hand. Thus the children were divided into two groups: high (n=27) and low (n=16) 18-month AHA score. In the MACS model the children were divided into three groups depending on their MACS level. To evaluate which random effects to include in the models and to compare models, the Bayesian Information Criterion was used (see Singer & Willet 2003). Limit and rate for each group were calculated as well as the 95% CI for these parameters. Limits and rates were compared between groups within each model. To enhance interpretation of the rate parameter, 'age-90' was calculated. Age-90 is the age at which 90% of the child’s limit is reached, and it was introduced by Rosenbaum et al. (2002). To calculate age-90 in this study, Formula 5 was rewritten as:
Age-90 = ln((limit-start)/limit-0.9*limit)/rate \tag{6}

The average and 95% confidence limits for age-90 were calculated using a bootstrap technique, non-parametric with 3,000 resamples (see example in Genovese 2008). The analyses and graphs were performed using the statistical software R, Version 2.8.0 (nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-89, and boot: Bootstrap R (S-Plus) Functions. R package version 1.2-34.)
ETHICAL CONSIDERATIONS

All studies in this thesis have been approved by the Ethics Research Committee at the Karolinska Hospital or the Regional Ethics Board of Karolinska Institutet in Stockholm, Sweden. Study III which involved data from Australia and the Netherlands was also approved by the Human Research Ethics Committee of Southern Health in Melbourne, Australia and the Commisie Mensgebonden Onderzoek Regio Arnhem-Nijmegen in the Netherlands. All children and families gave informed consent to participate in the studies.

The author of this thesis is a shareholder in the company Handfast AB (Inc.) which supplies training courses and test materials for the AHA. This company is supported and partially owned by Karolinska Institutet Innovations. The involvement with this company may be regarded as a conflict of interest. However, the company has not been involved in or funded any part of the research. Moreover, it does not own any of the data included in the studies in this thesis.
RESULTS
VALIDITY OF THE AHA

Study I

The internal scale evaluation showed that 20 of the 22 items (91%) fit the Rasch model assumptions. The items that misfit were *moves upper arm* and *moves forearm*. To investigate and reduce this misfit, stepwise item reduction was conducted, whereby the misfitting items were excluded. In the resulting 20-item scale item fit was 95%. These results indicated that the two misfitting items needed further scrutiny.

Person response evaluation showed that 98% of person measures fit the Rasch model assumptions. In the 20-item scale this figure was 97%. These were both within acceptable levels. The targeting of item difficulties in relation to person abilities was good. The difference between mean person measures and mean item measures was 0.16 logits. The targeting is illustrated in Figure 4. In the figure measures for the three thresholds between scale steps for each item are shown.

Supplementary analysis

The PCA of the 22-item scale showed that 82.6% of the variance could be explained by the dimension of the AHA scale, and 1.9% of the variance was in the largest secondary dimension. In a PCA for the 20-item scale, 84.6% of the variance could be explained by the AHA dimension and 1.8% by the next largest dimension. Thus deletion of the two items does not have a large impact on dimensionality in this analysis. This PCA-analysis indicates that the 22-items in the AHA constitute a construct that demonstrates unidimensionality to a satisfactory extent and that the second largest dimension in the scale is small.

The inspection of threshold measures for each item showed that 20 items had ordered thresholds. This indicates that in these 20 items the rating scale was used as expected and that all scale steps represented performance that was observed in the measured children. Two items had rating scale categories that were rarely used (*approaches objects*, score 3 had six responses and *puts down*, score 2 had nine responses) which led to disordered thresholds. Performance represented by these scale steps was rarely observed in the children, which may indicate a need to revise these items.

The DIF plot between the diagnostic groups unilateral CP and (Figure 5) showed that eight items fell outside the 95% confidence intervals drawn in the plot indicating DIF, including the two misfitting items: *moves upper arm* and *moves forearm*. The items *moves upper arm*, *moves forearm*, *reaches* and *moves fingers* were more difficult for the children with OBPP, whereas the items *releases*, *readjusts grip*, *manipulates* and *puts down* were more difficult for the children with unilateral CP. When comparing person measures obtained from the 22-item scale and from a scale with the eight DIF items removed, there was a 1.05 logit difference in mean person measure for the whole group (n=409). This indicates that the scale functions differently for the two diagnostic groups. The comparison between different age groups showed a clear DIF in five items.
### Results

<table>
<thead>
<tr>
<th>Measure line</th>
<th>Person ability</th>
<th>Item difficulty</th>
<th>score 1-2</th>
<th>score 2-3</th>
<th>score 3-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>more able</td>
<td>harder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each '.' is 1 person.
Each '#' is 3 persons.

**Items:**

1. Approaches objects
2. Holds
3. Stabilizes by weight or support
4. Proceeds
5. Changes strategies
6. Coordinates
7. Initiates use
8. Moves upper arm
9. Moves fingers
10. Orient objects
11. Stabilizes by grip
12. Releases
13. Flow in bimanual task performance
14. Chooses assisting hand when closer to objects
15. Calibrates
16. Reaches
17. Varies type of grasp
18. Readjusts grip
19. Grasps
20. Moves forearm
21. Manipulates
22. Puts down

---

Figure 4. Targeting of difficulty of items in the Assisting Hand Assessment to person ability in the 409 assessments in Study I.
Results

(Figure 6). The items moves fingers and moves upper arm were more difficult for the older children and the items changes strategies, calibrates and manipulates were more difficult for the smaller children. All DIF items fell close to the 95% confidence intervals, which were very narrow (Figure 6). When comparing mean person ability measures from the full 22-item scale and from a scale with these five items deleted, the difference was 0.07 logits. This indicates that the influence of DIF by age group does not affect person ability measures produced, and is thus not a threat to validity.

Figure 5. Plot of item measures of the AHA obtained from children with unilateral CP vs OBPP in the Differential Item Functioning analysis.

Figure 6. Plot of item measures of the AHA obtained from younger (≤60 months) and older children in the Differential Item Functioning analysis.
Results

RELIABILITY OF THE AHA

Study I showed that person separation reliability was 0.97 in the 22-item scale and for the reduced scale with 20 items 0.96. The person separation index was 5.42 in the scale with 22 items and 5.20 in the 20-item scale. Both of these translate to seven distinctly different ability strata.

Rater reliability (Study II)

Between raters the ICC for sum scores was 0.98 in the 2-rater design and 0.97 in the 20-rater design. Within raters the ICC was 0.99, for CIs see Table VII. The ICC for individual items varied considerably (Table VII). The SEM between raters was 1.5, producing an error interval of ±3 raw scores. For intraraters the SEM was 1.2, producing an interval of ±2.4 raw scores.

The comparison between raters’ experience and scoring showed no correlations. This indicates that all raters had sufficient training to conduct reliable AHA assessments.

Retest and alternate forms reliability (Study III)

Retest ICC for the Small Kids version was 0.99 and for the School Kids version 0.98. Alternate forms ICC:s were 0.99 between Small- and School Kids versions and 0.98 between the two board games in the School Kids AHA (for CI see Table VII). Retest and alternate forms ICC:s for individual items varied from -0.03 to 1.00 (Table VII).

The paired t-test showed no systematic differences between test and retest or between versions. For individual items with an ICC lower than 0.70, which according to Polit and Hungler (1995) is the lowest acceptable ICC for group level comparisons, the comparison between test times and versions showed no systematic differences.

The SEM for retest was 0.35 logits for the Small Kids version and 0.28 logits for School Kids (see Table VI for raw scores). This produced SDDs of 0.97 and 0.76 logits respectively.

Table VI. Standard Error of Measurement and Smallest Detectable Difference for different reliability evaluations.

<table>
<thead>
<tr>
<th></th>
<th>SEM</th>
<th>SDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interrater</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Intrarater</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Kids</td>
<td>1.4</td>
<td>3.9</td>
</tr>
<tr>
<td>School Kids</td>
<td>1.3</td>
<td>3.7</td>
</tr>
<tr>
<td>Alternate forms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small vs. School Kids</td>
<td>1.2</td>
<td>3.2</td>
</tr>
<tr>
<td>Alien-Fortress</td>
<td>1.3</td>
<td>3.7</td>
</tr>
</tbody>
</table>

*aCalculated from data in the Alternate forms: Alien-Fortress evaluation.

The Bland-Altman plots showed that assessments of alternate forms were all within the limits of agreement. Common person and item linking plots showed that all item difficulties and person measures were within the 95% confidence interval lines. This
indicated that test items and the tests as a whole were of equal difficulty between versions, and that different versions measure the same construct.

Table VII. Intraclass correlation coefficients for reliability of the AHA in different designs.

<table>
<thead>
<tr>
<th></th>
<th>Inter-rater: 2 raters</th>
<th>Inter-rater: 20 raters</th>
<th>Intra-rater</th>
<th>Retest Small Kids</th>
<th>Alternate forms Small vs School Kids</th>
<th>Alternate forms &amp; Retest School Kids</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sum score</strong></td>
<td>.98</td>
<td>.97</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
</tr>
<tr>
<td>95% Confidence Interval</td>
<td>.95-.99</td>
<td>.92-.99</td>
<td>.98-.99</td>
<td>.97-.99</td>
<td>.98-.99</td>
<td>.96-.99</td>
</tr>
<tr>
<td><strong>Items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>General use items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approaches objects</td>
<td>1.00</td>
<td>.91</td>
<td>.95</td>
<td>.80</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Initiates use</td>
<td>.77</td>
<td>.89</td>
<td>1.00</td>
<td>.75</td>
<td>.83</td>
<td>.59</td>
</tr>
<tr>
<td>Chooses assisting hand when closer to objects</td>
<td>.41</td>
<td>.81</td>
<td>.89</td>
<td>NA</td>
<td>1.00</td>
<td>-.03</td>
</tr>
<tr>
<td><strong>Arm use items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stabilizes by weight or support</td>
<td>.93</td>
<td>.85</td>
<td>.95</td>
<td>.98</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Reaches</td>
<td>.78</td>
<td>.86</td>
<td>.94</td>
<td>.75</td>
<td>.95</td>
<td>.82</td>
</tr>
<tr>
<td>Moves upper arm</td>
<td>.53</td>
<td>.70</td>
<td>.94</td>
<td>.12</td>
<td>.88</td>
<td>.92</td>
</tr>
<tr>
<td>Moves forearm</td>
<td>.35</td>
<td>.79</td>
<td>.88</td>
<td>.91</td>
<td>1.00</td>
<td>.92</td>
</tr>
<tr>
<td><strong>Grasp - release items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grasps</td>
<td>.79</td>
<td>.90</td>
<td>.91</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Holds</td>
<td>.94</td>
<td>.94</td>
<td>.98</td>
<td>.94</td>
<td>.83</td>
<td>1.00</td>
</tr>
<tr>
<td>Stabilizes by grip</td>
<td>.86</td>
<td>.85</td>
<td>.97</td>
<td>.93</td>
<td>.94</td>
<td>.96</td>
</tr>
<tr>
<td>Readjusts grip</td>
<td>.77</td>
<td>.88</td>
<td>.88</td>
<td>.94</td>
<td>.93</td>
<td>.80</td>
</tr>
<tr>
<td>Varies type of grasp</td>
<td>.93</td>
<td>.89</td>
<td>.86</td>
<td>.86</td>
<td>.97</td>
<td>.75</td>
</tr>
<tr>
<td>Releases</td>
<td>.75</td>
<td>.94</td>
<td>.98</td>
<td>.94</td>
<td>.94</td>
<td>.89</td>
</tr>
<tr>
<td>Puts down</td>
<td>NA</td>
<td>.82</td>
<td>.92</td>
<td>.00</td>
<td>1.00</td>
<td>.65</td>
</tr>
<tr>
<td><strong>Fine motor adjustment items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moves fingers</td>
<td>.73</td>
<td>.77</td>
<td>.85</td>
<td>.82</td>
<td>.91</td>
<td>.90</td>
</tr>
<tr>
<td>Calibrates</td>
<td>.91</td>
<td>.84</td>
<td>.92</td>
<td>.91</td>
<td>.96</td>
<td>.90</td>
</tr>
<tr>
<td>Manipulates</td>
<td>.46</td>
<td>.80</td>
<td>.98</td>
<td>.79</td>
<td>.76</td>
<td>.59</td>
</tr>
<tr>
<td><strong>Coordination items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordinates</td>
<td>.74</td>
<td>.87</td>
<td>.94</td>
<td>.92</td>
<td>.93</td>
<td>.83</td>
</tr>
<tr>
<td>Orient objects</td>
<td>.69</td>
<td>.86</td>
<td>.95</td>
<td>.94</td>
<td>.94</td>
<td>.87</td>
</tr>
<tr>
<td><strong>Pace items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proceeds</td>
<td>.67</td>
<td>.77</td>
<td>.84</td>
<td>.80</td>
<td>.64</td>
<td>.80</td>
</tr>
<tr>
<td>Changes strategy</td>
<td>.76</td>
<td>.92</td>
<td>.89</td>
<td>.55</td>
<td>.57</td>
<td>.85</td>
</tr>
<tr>
<td>Flow in bimanual task performance</td>
<td>.73</td>
<td>.84</td>
<td>.95</td>
<td>.93</td>
<td>1.00</td>
<td>.92</td>
</tr>
</tbody>
</table>

NA=ICC could not be calculated, because no variance was present since all participants received the same score in compared assessments.

DEVELOPMENT OF ASSISTING HAND USE (STUDY IV)

Demographic data for the 43 children is reported in Table VIII. In all children an increase in AHA scores was seen over time. For three children this increase was smaller than the SDD, and for three other children the increase was followed by a significant decrease in AHA scores. There were considerable inter-individual differences in improvement as shown in Figure 7.

In the final 18-month AHA and MACS models of development the limit and rate were included as both random and fixed effects. The children who scored less than -3 logits
at 18 months (n=16, children 1-16 in Table VIII) displayed on average a larger increase in ability than the children in the high 18-month AHA group (n=27). There was a

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Hemiplegic side</th>
<th>GA</th>
<th>MACS level</th>
<th>Inclusion age (mo)</th>
<th>Time in study (mo)</th>
<th>No of assessments</th>
<th>% logits</th>
<th>AHA at inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>boy</td>
<td>left</td>
<td>36</td>
<td>III</td>
<td>22</td>
<td>72</td>
<td>7</td>
<td>2</td>
<td>-8.81</td>
</tr>
<tr>
<td>2</td>
<td>boy</td>
<td>right</td>
<td>31</td>
<td>III</td>
<td>18</td>
<td>86</td>
<td>9</td>
<td>3</td>
<td>-6.76</td>
</tr>
<tr>
<td>3</td>
<td>girl</td>
<td>left</td>
<td>39</td>
<td>II</td>
<td>19</td>
<td>68</td>
<td>6</td>
<td>5</td>
<td>-7.25</td>
</tr>
<tr>
<td>4</td>
<td>girl</td>
<td>right</td>
<td>40</td>
<td>III</td>
<td>21</td>
<td>86</td>
<td>4</td>
<td>6</td>
<td>-6.76</td>
</tr>
<tr>
<td>5</td>
<td>boy</td>
<td>left</td>
<td>na</td>
<td>III</td>
<td>21</td>
<td>37</td>
<td>6</td>
<td>6</td>
<td>-6.76</td>
</tr>
<tr>
<td>6</td>
<td>boy</td>
<td>right</td>
<td>40</td>
<td>III</td>
<td>49</td>
<td>54</td>
<td>6</td>
<td>6</td>
<td>-6.76</td>
</tr>
<tr>
<td>7</td>
<td>girl</td>
<td>left</td>
<td>34</td>
<td>III</td>
<td>25</td>
<td>35</td>
<td>5</td>
<td>8</td>
<td>-6.35</td>
</tr>
<tr>
<td>8</td>
<td>boy</td>
<td>right</td>
<td>37</td>
<td>III</td>
<td>23</td>
<td>47</td>
<td>6</td>
<td>8</td>
<td>-6.35</td>
</tr>
<tr>
<td>9</td>
<td>boy</td>
<td>left</td>
<td>31</td>
<td>II</td>
<td>18</td>
<td>42</td>
<td>7</td>
<td>9</td>
<td>-6.00</td>
</tr>
<tr>
<td>10</td>
<td>girl</td>
<td>right</td>
<td>39</td>
<td>II</td>
<td>18</td>
<td>63</td>
<td>8</td>
<td>11</td>
<td>-5.69</td>
</tr>
<tr>
<td>11</td>
<td>girl</td>
<td>right</td>
<td>40</td>
<td>II</td>
<td>33</td>
<td>38</td>
<td>7</td>
<td>21</td>
<td>-4.02</td>
</tr>
<tr>
<td>12</td>
<td>girl</td>
<td>left</td>
<td>na</td>
<td>III</td>
<td>18</td>
<td>66</td>
<td>7</td>
<td>24</td>
<td>-6.00</td>
</tr>
<tr>
<td>13</td>
<td>boy</td>
<td>right</td>
<td>40</td>
<td>III</td>
<td>44</td>
<td>57</td>
<td>8</td>
<td>24</td>
<td>-3.55</td>
</tr>
<tr>
<td>14</td>
<td>boy</td>
<td>right</td>
<td>40</td>
<td>II</td>
<td>19</td>
<td>75</td>
<td>11</td>
<td>24</td>
<td>-3.55</td>
</tr>
<tr>
<td>15</td>
<td>boy</td>
<td>left</td>
<td>39</td>
<td>III</td>
<td>35</td>
<td>40</td>
<td>7</td>
<td>26</td>
<td>-3.32</td>
</tr>
<tr>
<td>16</td>
<td>boy</td>
<td>right</td>
<td>39</td>
<td>III</td>
<td>46</td>
<td>48</td>
<td>4</td>
<td>26</td>
<td>-3.32</td>
</tr>
<tr>
<td>17</td>
<td>girl</td>
<td>right</td>
<td>40</td>
<td>III</td>
<td>25</td>
<td>36</td>
<td>5</td>
<td>33</td>
<td>-2.23</td>
</tr>
<tr>
<td>18</td>
<td>girl</td>
<td>left</td>
<td>29</td>
<td>II</td>
<td>19</td>
<td>60</td>
<td>8</td>
<td>38</td>
<td>-1.59</td>
</tr>
<tr>
<td>19</td>
<td>boy</td>
<td>right</td>
<td>37</td>
<td>II</td>
<td>23</td>
<td>57</td>
<td>7</td>
<td>38</td>
<td>-1.59</td>
</tr>
<tr>
<td>20</td>
<td>boy</td>
<td>right</td>
<td>39</td>
<td>II</td>
<td>27</td>
<td>38</td>
<td>7</td>
<td>38</td>
<td>-1.59</td>
</tr>
<tr>
<td>21</td>
<td>boy</td>
<td>left</td>
<td>38</td>
<td>II</td>
<td>46</td>
<td>45</td>
<td>5</td>
<td>41</td>
<td>-1.16</td>
</tr>
<tr>
<td>22</td>
<td>boy</td>
<td>right</td>
<td>38</td>
<td>II</td>
<td>19</td>
<td>68</td>
<td>11</td>
<td>41</td>
<td>-1.16</td>
</tr>
<tr>
<td>23</td>
<td>boy</td>
<td>right</td>
<td>30</td>
<td>II</td>
<td>51</td>
<td>38</td>
<td>6</td>
<td>42</td>
<td>-0.93</td>
</tr>
<tr>
<td>24</td>
<td>girl</td>
<td>left</td>
<td>39</td>
<td>II</td>
<td>38</td>
<td>61</td>
<td>7</td>
<td>42</td>
<td>-0.93</td>
</tr>
<tr>
<td>25</td>
<td>girl</td>
<td>left</td>
<td>27</td>
<td>II</td>
<td>36</td>
<td>35</td>
<td>4</td>
<td>42</td>
<td>-0.93</td>
</tr>
<tr>
<td>26</td>
<td>girl</td>
<td>left</td>
<td>37</td>
<td>II</td>
<td>36</td>
<td>57</td>
<td>7</td>
<td>42</td>
<td>-0.93</td>
</tr>
<tr>
<td>27</td>
<td>girl</td>
<td>left</td>
<td>42</td>
<td>II</td>
<td>21</td>
<td>76</td>
<td>8</td>
<td>44</td>
<td>-0.70</td>
</tr>
<tr>
<td>28</td>
<td>girl</td>
<td>right</td>
<td>37</td>
<td>II</td>
<td>28</td>
<td>40</td>
<td>8</td>
<td>45</td>
<td>-0.47</td>
</tr>
<tr>
<td>29</td>
<td>boy</td>
<td>right</td>
<td>32</td>
<td>II</td>
<td>29</td>
<td>51</td>
<td>6</td>
<td>48</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>boy</td>
<td>right</td>
<td>39</td>
<td>II</td>
<td>24</td>
<td>82</td>
<td>7</td>
<td>48</td>
<td>0.01</td>
</tr>
<tr>
<td>31</td>
<td>girl</td>
<td>left</td>
<td>30</td>
<td>II</td>
<td>29</td>
<td>66</td>
<td>6</td>
<td>50</td>
<td>0.25</td>
</tr>
<tr>
<td>32</td>
<td>girl</td>
<td>left</td>
<td>43</td>
<td>II</td>
<td>22</td>
<td>58</td>
<td>7</td>
<td>50</td>
<td>0.25</td>
</tr>
<tr>
<td>33</td>
<td>girl</td>
<td>left</td>
<td>32</td>
<td>II</td>
<td>63</td>
<td>26</td>
<td>3</td>
<td>52</td>
<td>0.49</td>
</tr>
<tr>
<td>34</td>
<td>girl</td>
<td>right</td>
<td>42</td>
<td>I</td>
<td>29</td>
<td>64</td>
<td>8</td>
<td>53</td>
<td>0.73</td>
</tr>
<tr>
<td>35</td>
<td>girl</td>
<td>right</td>
<td>40</td>
<td>II</td>
<td>45</td>
<td>58</td>
<td>8</td>
<td>53</td>
<td>0.73</td>
</tr>
<tr>
<td>36</td>
<td>girl</td>
<td>left</td>
<td>38</td>
<td>II</td>
<td>30</td>
<td>64</td>
<td>7</td>
<td>55</td>
<td>0.96</td>
</tr>
<tr>
<td>37</td>
<td>girl</td>
<td>right</td>
<td>29</td>
<td>I</td>
<td>35</td>
<td>68</td>
<td>6</td>
<td>55</td>
<td>0.96</td>
</tr>
<tr>
<td>38</td>
<td>girl</td>
<td>right</td>
<td>40</td>
<td>II</td>
<td>60</td>
<td>35</td>
<td>4</td>
<td>56</td>
<td>1.20</td>
</tr>
<tr>
<td>39</td>
<td>boy</td>
<td>left</td>
<td>37</td>
<td>I</td>
<td>60</td>
<td>33</td>
<td>3</td>
<td>59</td>
<td>1.65</td>
</tr>
<tr>
<td>40</td>
<td>boy</td>
<td>right</td>
<td>31</td>
<td>I</td>
<td>39</td>
<td>49</td>
<td>6</td>
<td>64</td>
<td>2.31</td>
</tr>
<tr>
<td>41</td>
<td>boy</td>
<td>right</td>
<td>30</td>
<td>I</td>
<td>23</td>
<td>31</td>
<td>7</td>
<td>67</td>
<td>2.72</td>
</tr>
<tr>
<td>42</td>
<td>boy</td>
<td>right</td>
<td>42</td>
<td>I</td>
<td>37</td>
<td>48</td>
<td>7</td>
<td>74</td>
<td>3.61</td>
</tr>
<tr>
<td>43</td>
<td>boy</td>
<td>right</td>
<td>32</td>
<td>I</td>
<td>64</td>
<td>38</td>
<td>4</td>
<td>92</td>
<td>5.66</td>
</tr>
</tbody>
</table>
Results

A significant difference in the average limit AHA scores attained, -0.24 logits for the low 18-month AHA group and 2.14 for the group with higher ability. The more able group reached 90% of their maximum level at a lower age (3 years) than the children in the low 18-month AHA-score group who reached their age-90 at 7 years 3 months of age (Figure 8).

When dividing the children according to their MACS level a similar trend was seen (Figure 9). The children in MACS level I (n=7) attained a high maximum limit (4.42 logits) at a young age (90% at age 3 years and 9 months). The children in MACS level...
Results

II (n=25) reached a significantly lower limit (1.32 logits), but they attained their age-90 at an age about as low as the MACS I group (4 years). The children in MACS level III (n=11) displayed an average limit that was lower than the MACS II group, and they took significantly longer time to attain their limit. The estimated average age-90 was 9 years 5 months, which was outside the age range measured in this study.

Figure 9. Observed and predicted development for the three levels of Manual Ability Classification System. Vertical lines represent age-90 (solid lines) and its 95% confidence interval (dotted lines).
DISCUSSION

PSYCHOMETRIC PROPERTIES OF THE AHA

The studies in this thesis support the psychometric properties and the usefulness of the AHA in various ways. The promising results from the paper describing the initial development of the AHA in 2003 (Krumlinde-Sundholm & Eliasson) have been expanded upon and confirmed in this thesis. Strong evidence is presented of its validity based on internal structure, response processes and relationship to other variables is presented. This indicates that that strong confidence can be placed in the interpretation of AHA measures for children with unilateral CP and OBPP. Furthermore, its reliability was in all aspects found to be excellent; AHA measures are stable across test sessions and versions and the agreement among and within trained raters is very high. Results from Study IV showed that the AHA can be used to measure change in individuals over time.

As regards evidence based on internal structure the two different analyses of unidimensionality in Study I and in the supplementary analysis gave interesting results. It is evident that inspection of goodness-of-fit statistics and the PCA measure different aspects of the detection of multidimensionality. Linacre (1996) compares these two types of analysis. He states that the goodness-of-fit statistics merely evaluate one item at a time and that the detection of multidimensionality requires examination of the relationship between the items. Therefore the PCA should carry more weight in the dimensionality analysis. Smith (2002) further reports that where there are two or more dimensions in a scale with approximately equal number of items in each dimension the PCA is better able to detect multidimensionality. When the majority of items contribute to one component, the fit statistics are better able to detect departures from unidimensionality. Since the PCA indicated unidimensionality, it appears that the AHA does not contain two or more equal sets of items representing separate dimensions, but rather one main dimension. On the other hand there were misfitting items in the goodness-of-fit analysis that indicate a departure from the scale dimension by two of the items. In the discussion in Study I, we hypothesized that the misfit would be due to the differences between the two diagnostic groups assessed in their ability to move the arm. For children with a unilateral CP, the impairment has the largest impact on hand movements. Whereas for children with OBPP, especially Erb’s palsy the impairment primarily affects shoulder and arm movements. This hypothesis was confirmed by the DIF analysis in this summary. My conclusion is that the AHA to a satisfactory extent is unidimensional, and that the misfitting items are not a threat to unidimensionality, but rather a reflection of DIF. To remedy this DIF, there is a need to create separate scales for children with unilateral CP and OBPP. It is likely that the present item hierarchy is based more on the abilities of children with unilateral CP since they had more representation in Study I. Therefore, larger changes in the item hierarchy for children with OBPP than with unilateral CP are expected. The unidimensionality of these separate scales must be confirmed in a later study.
Further evidence of internal structure involves the rating scale structure. By using the partial credit model it was made possible to examine the threshold measures of each item. This was informative and showed good results for 20 of the items. The disordered thresholds of the items *approaches objects* and *puts down* were expected. With both items, the scale steps that are infrequently used have surrounding thresholds that are close to each other, indicating narrow steps (see Figure 4). Both these items need to be further evaluated and possibly changed.

Strong evidence of validity based on response processes were presented in the analyses person response validity and targeting of item difficulty to person ability. Evidence for validity based on the relationship to other variables in terms of DIF between the diagnostic groups was discussed above. As regards the DIF-analysis between age groups a small age DIF was found in five items. This indicates that there are minor differences in the item hierarchy between the age groups. From what we know of typical development these differences are quite expected and were also anticipated in the discussion in Study I. For example it is known that small children have not yet developed calibration of grip force and in-hand manipulation (Forssberg et al. 1992; Newell & McDonald 1997). These are among the more difficult items in the AHA scale and rarely develops fully in children with unilateral CP or OBPP. The influence of both cognitive development and experience would give an advantage to older children in the item *changes strategies*. The evaluation of person measures showed that invariance remains despite presence of DIF items, indicating that this DIF is not a threat to a valid interpretation of AHA results.

Evidence of validity in relationship to other variables includes the common person linking analysis in Study III. The result that all tested versions measured the same construct is good, but not surprising since the same items are used in all versions. It would be interesting to compare the AHA to another test with a similar purpose. The reason for developing the AHA was that there was no test available for measuring assisting hand use. Therefore this comparison has not been possible.

In this thesis, the overall reliability of AHA measures has been shown to be excellent. The internal consistency expressed as the person separation reliability was found to be high. This indicates homogeneity among items. The high person separation index that gives a separation into seven distinctly different ability strata indicates that the AHA has a long scale. This is one of the strengths of the AHA; it can give meaningful measures both to children that do not use their affected hand at all, as well as to children with a disability that is only visible to the observer after careful observation.

Reliability coefficients for stability and equivalence are very high. This is an indication that, for example, raters can distinguish well between individuals of different ability. These high coefficients should not, however, lead to the conclusion that the error surrounding an observed score is negligible. Comparing the sources of error in the AHA, Table VI can be studied. In the case of AHA all SEMs are influenced by rater error. For a single assessment the AHA score ±2 SEM_{intrater} is the score interval within which the true score lies with 95% certainty. The SEM:s for alternate forms between
Discussion

Small and School Kids and the retest SEM:s for both Small and School Kids are practically the same. This would indicate that changing between Small and School Kids versions has no or minimal effect on the scores. The comparison between retest (Small Kids) and intrarater is interesting. In both cases the same rater is rating the sessions. The difference lies in the repeated testing of the children which adds only 0.2 raw scores to the SEM. This would imply that the majority of error would be due to the rater. A conclusion of this is that the AHA seems to be stable in the sense that children’s performances are very similar from one testing to the next.

When comparing scores between two assessments the retest SDD is applicable. A change from one assessment to the next that is smaller than the SDD may not be a real change, but due to chance, day-to-day fluctuations within the child, or more likely due to rater. A change that in AHA scores that exceeds the SDD is considered a real and statistically significant change. To evaluate whether a test is sensitive to change, the SDD can be compared to the amount of change seen from intervention or development (MacDermid & Stratford 2004). In the description of individual development in Study IV, we can see that over a longer time the vast majority of children did make a change of AHA scores that exceeded the SDD. Thus we have seen that long-term change in individuals is larger than the SDD. In Eliasson et al. (2005) significant changes on a group level were reported after a treatment period. On an individual level, however, measures before and after treatment showed a change that was higher than the SDD in 10 of the 21 children in the study group, but actually all children but one showed some positive change. The differences seen in the comparison group (n=20) were smaller, only three children showed improvements exceeding the SDD, and the rest showed changes less than 0.5 logits (about 2 raw scores). These small changes in the comparison group within the time period of two months indicate that the AHA-meaures are stable when no change is expected and even fluctuating less than the SDD.

DEVELOPMENT OF ASSISTING HAND USE

The results from Study IV, where all children improve over time, are interesting and encouraging. The fact that these children had access to continuous evaluation and intervention through contact with habilitation centers is likely to have influenced the results. Unfortunately, we can not evaluate the extent of this influence from available data. This would require a similar sample without access to interventions. As can be seen in Figure 7 there are deviations in the curves which often, but far from always, are timed at periods of intensive intervention. My speculation is that the observed curves would have had smaller deviations without this intervention. It is also likely that the different interventions had a positive impact on long term development.

The children in MACS level III continued to slowly improve their assisting hand use throughout the study as indicated by the high age-90 (9 years, 5 months). The non-linear model does allow that conclusions are drawn outside of the measured age range. Therefore reporting this result is not a violation of the method. On the other hand conclusions outside of what is observed in this study are uncertain for all MACS...
groups. This is because the groups were small and development varied within the groups.

Studying the results from Study IV raises additional questions: At what age will the children in MACS level III reach their limit? Will some children deteriorate at a higher age as with gross motor function (Hanna, Rosenbaum et al. in press)? With the development of AHA for teenagers and adults it will be possible to continue to follow these children. Other questions that are raised regard the origin of the large differences in development between individuals. These questions will be addressed in a later study where these development curves will be further analyzed to find possible predictors.

**METHODOLOGICAL CONSIDERATIONS**

The choice of sample size in Study I was based on the reasoning of Linacre (1994). His recommendations for sample sizes for definitive item calibrations were 250 to 20*test length. For AHA this would add up to 440 assessments. Moreover, Tennant identified 400-500 as the ideal number of assessments to include in a Rasch analysis for a stable evaluation of a test (Alan Tennant, personal communication 28 November 2006). For the reliability studies (II and III), the sample size was based on calculations by Walter, Eliasziw and Donner (1998).

In Study IV a larger, randomly selected study group would have improved generalizability of results. The size of the present group gave large confidence intervals on comparisons between subgroups. There were however significant differences between some groups, indicating that development was quite different between these groups. This signifies that the aim of the study was still fulfilled. Because this is the first longitudinal study of assisting hand use results are highly relevant. The convenience sample used in Studies I-III has minor influence on the generalizability of the results. For these studies the sample should be regarded as representative for the groups of children that the test is intended for.

The choice of a non-linear model for the analysis of data in Study IV was not straightforward. In a preliminary analysis a linear random coefficient model was used to analyze the data. In this analysis, the development curves took the shape of a second order polynomial (quadratic change), which by default indicates a decreased AHA score in higher ages. As noted in the results section, this was not the case from empirical data for 40 of the 43 children. Thus, the linear model did not correspond either to data or to our theoretical understanding of development. We therefore chose to use a non-linear model although this entailed to some complications. The first complication was that it requires a large data set, preferably larger than the one present in Study IV. This problem was overlooked since it was actually possible to perform the calculations. The second complication was that it was not a general model that is available in any statistical software. A mathematical function that corresponds to the theoretical assumptions about development must be applied. This requires profound knowledge in both statistics and software programming. This complication was resolved by incorporating a professional statistician as a co-worker.
The decision to assume that all children have a minimum AHA score at age 0 months in the analysis of Study IV is correct in one sense; both typically developing children and children with CP are unable to handle objects bimanually at birth and thus would have a minimum score on the AHA. On the other hand, this reasoning is flawed, because it is not possible to measure children validly with the AHA at birth. Therefore, this assumption is merely theoretical. Moreover, there is no knowledge about the course of development from birth to age 18 months. We therefore used the minimum score as the start level in the analysis, but intentionally started drawing the average development curves from age 18 months.

The way in which we divided children into two groups by their AHA score at 18 months in order to illustrate and contrast their development may seem arbitrary. We found, however, that the difference in performance between the children with AHA scores above and below -3 logits was so marked that this division was justified.

The choice to use the SDD as a criterion for change in Study IV can be debated since we did not have the same procedure for scoring as in the retest study (Study III). It is most likely that the SDD will be smaller when using the procedure that was used in Study IV. It would be interesting to conduct a study with the purpose of estimating the SDD for this procedure. The way in which children were scored in Study IV, by rechecking scores, has probably revealed if children were mistakenly scored as having changed when there was, in reality, no change in performance. If the opposite mistake was made, the child’s performance had, in fact, changed but this was not captured in the assessment, this was not revealed by the subsequent recheck. Thus, it is possible that some change in the children in Study IV was not described due to intrarater error.

FUTURE RESEARCH

This thesis answered many questions about the psychometric properties of the AHA and development of assisting hand use, while some new questions were also raised.

One topic for future research is to continue to develop the AHA scale. This includes creating separate AHA scales for children with unilateral CP and children with OBPP as well as revising the items approaches objects and puts down. The advantage from interval level measures obtained from a Rasch analysis needs to be made easier to interpret and accessible for AHA users. This can be done by transforming the logit measures to e.g., a 0-100 score. This is what has been done with the Gross Motor Function Measure 66-item version (Russell et al. 2000) and the scaled scores for the Pediatric Evaluation of Disability Inventory (PEDI, Haley et al. 1992). This would make the logit measures, which are preferable to raw score, more easily interpretable and thus be beneficial both for clinical use and research.

As mentioned earlier the development of AHA for use with other diagnostic and age groups is ongoing. This includes children with acquired brain injury, with upper limb reduction deficiency, with bilateral CP and children that use a prosthesis. This even includes children younger than 18 months, adolescents and adults.
Study IV has two immediate follow-up opportunities. The first is to further evaluate the existing data with the addition of other variables such as type of brain lesion to find possible predictors of different development. The second is to continue to follow these children into adolescence and even extend the group to include more children predominantly in MACS level I.
CONCLUSIONS

This thesis reports evidence that the AHA validly measures assisting hand use in children with unilateral CP and OBPP. The AHA has a sound internal structure that is unidimensional, excellent response processes as well as good evidence of validity in relationship to other variables. This validity evidence indicates that confidence can be placed in the interpretation of AHA measures.

The AHA has excellent ability to distinguish between individuals of different ability. This is reflected in the high reliability coefficients for internal consistency, stability and equivalence as well as a high person separation index. When measuring change with the AHA the smallest detectable change is less than 4 raw scores or just under 1 logit.

Development curves for children with unilateral CP between the ages of 18 months and 8 years using the AHA were presented. Results show a significantly different course of development based on the AHA score at 18 months. Results further show that development of assisting hand use was different for children in different MACS levels. These development curves can be used when discussing prognosis of assisting hand development for children with unilateral CP.

Syftet med denna avhandling är att utvärdera AHA:s psykometriska egenskaper för det utökade åldersspannet samt att beskriva hur användningen av den påverkade handen utvecklas hos barn med unilateral CP.


Reliabiliteten hos AHA har prövats både inom och mellan bedömare samt för test–retest och mellan olika versioner av AHA. Resultaten från alla dessa prövningar visar höga reliabilitetskoefficienter (Intraclass Correlation Coefficient), som varierar mellan 0.97 och 0.99. Test–retestprövningen visar också att den minsta förändring som kan mätas med AHA är 4 poäng, vilket motsvarar mindre än 6 procent av skalan. Resultaten visar vidare att alla AHA-versioner som prövats mot varandra ger samma poäng och mäter samma fenomen. Detta gäller alltså såväl småbarns-AHA som de två spelen i skolbarns-AHA.

Hos 43 barn med unilateral CP har hjälphandsfunktionens utveckling studerats för åldrarna 18 månader till 8 år. Barnen testades regelbundet med AHA under i genomsnitt 4.5 år. Alla barn i studien ökade sin förmåga att använda den hemiplegiska handen över tid. En icke-linjär "mixed model" används för att skapa separat
Sammanfattning på svenska


Sammanfattningsvis visar studierna i denna avhandling att AHA kan användas för att göra giltiga och pålitliga bedömningar av hur effektivt barn med unilateral CP och brachial plexusskada använder sin påverkade hand i bimanuella aktiviteter. AHA kan också användas för att följa hur denna typ av handfunktion utvecklas.
ACKNOWLEDGEMENTS

This thesis represents a learning experience that has been both challenging and enjoyable and which I am deeply grateful for. The thesis also represents teamwork in more than one sense. Some people have been deeply involved in the research while many others have been important supporters in other ways. I want to express my sincere gratitude to all those who have contributed to this thesis, especially to:

My two supervisors for taking me into your team and trusting me to work with you on the AHA. You have been extraordinarily generous with your time, viewpoints and material. I highly value your support and friendship and I look forward to more challenging work together in the future.

Professor Ann-Christin Eliasson, my main supervisor; I really appreciate your true enthusiasm for research and your passionate drive to find new ways to make life better for “our” children. You have shown me that research is both exciting and hard work, but seldom boring. Most of all, I want to thank you for being a great inspiration every time we meet and for, when needed, keeping my eyes on the goal and discarding irrelevant details that I tend to get occupied with.

Lena Krumlinde Sundholm, my supervisor and closest collaborator during these years. You have spent so many hours guiding and mentoring me and have taught me so much; from how to manage statistical software to how to teach about measurement and AHA. The way you keep digging for more knowledge and your bright new ideas are an inspiration. More than anything I want to thank you for challenging my thinking with always one more question of why or how and for your massive support and belief in me.

My co-authors; Anders Kottorp for sharing knowledge in Study I, Pauline Aarts and Brian Hoare for invaluable help with Study III and to Jakob Bergström for enthusiasm and expertise on mixed models in Study IV.

Professor Hans Forssberg for being a patron and important supporter of the paramedical research at our unit.

The children that participated in these studies and their families. Thank you for generously giving your time and energy for the possibility of new knowledge to be created. Meeting you has been a privilege and you will always be special to me.

The Occupational Therapists at Habilitation Centres in Stockholm and other regions in Sweden for your interest in this research and the invaluable help with data collection. Even though your schedules are always full of demanding duties, you have helped in any way you could.

My colleagues in the research group at the Neuropediatric unit, Department of Woman and Child Health. Our discussions over the years have been formative and I appreciate...
the open atmosphere at our seminars, allowing for any question. A special thanks to Annika Sköld, Ann-Marie Öhrvall, Linda Holmström Marie Berg and Kicki Löwing for sharing struggles and your knowledge and having fun together at conferences. I also wish to thank our administrator None-Marie Kemp and computer-whiz Mikael Reimeringer, for always being friendly and quick to help.

My colleagues at the former Occupational Therapy Department at Linköping University where I had my workplace during the first years. Thank you to Chris Henriksson for opening the door for me into this research environment and also to Gunilla Liedberg, Elin Ekbladh, Mathilda Björk and Jan Sandqvist for sharing many laughs and discussions on life and methodology during lunches and seminars. Anette Kjellberg and Lena Haglund were the ones who first made me interested in research and encouraged me to take on this journey.

My colleagues at the Centre for Rehabilitation Research in Örebro where I had my workplace during this last hectic year. Thank you for including me in your fellowship, allowing me space at seminars and being patient with my poor socializing during these last few months. A special thank you to Lars-Olov Lundqvist and Birgitta Sjökvist-Nätterlund for welcoming me and to Liselotte Hermansson for your encouragement and support and your constructive criticism on manuscripts.

My parents Siv and Bertil and in-laws Linnea and Per for being so supportive of our family and my work. Thank you for all of the invaluable practical help during these years. My brother Henrik and family for good company and also “siblings-in-law” Anders, Sussi and Nina and their families for friendship and good times whenever we meet.

All dear friends in Johanneskyrkan in Linköping and Mötesplatsen in Örebro for belonging together and sharing life with us.

Niklas, my husband and best friend. Thank you for your unconditional love, support, patience and encouragement during these years. You bring color and vision to my life and provide fantastic food at all times. Jonathan and Esther, I’m so proud of being your mom. You bring the sun out for me.

The Health Care Sciences Postgraduate School at Karolinska Institutet for funding these doctoral studies. I’m also very thankful for the excellent courses, moral support and nice dinners you provided.
REFERENCES


References


References


Ref Type: Computer Program


