Interrater and Intrarater Reliability of the Assisting Hand Assessment

Marie Holmefur, Lena Krumlinde-Sundholm, Ann-Christin Eliasson

KEY WORDS
• hand
• pediatric
• reliability

OBJECTIVE. The aim of this study was to evaluate interrater and intrarater reliability for the Assisting Hand Assessment.

METHOD. For interrater reliability, two designs were used: 2 occupational therapists rated the same 18 children, and 20 occupational therapists rated the same 8 children. For intrarater reliability, 20 raters each rated one child twice. Both English and Swedish versions of the instrument were used. Intraclass correlation coefficients (ICCs) and standard error of measurement (SEM) were calculated.

RESULTS. ICCs for the sum score for the interrater were 0.98 (two raters) and 0.97 (20 raters) and for the intrarater 0.99. SEM was 1.5 for interrater and 1.2 for the intrarater study, which gave an error interval of ± 3 raw scores for interrater and ± 2.4 raw scores for intrarater.

CONCLUSION. This study shows excellent interrater and intrarater reliability for sum scores.


Introduction

Hand skills are vital for the performance of activities in everyday life. The coordinated use of two hands is the most efficient way to conduct most activities. Impairments of the upper limbs imply complications in almost all areas of human occupations (Exner, 2001; Sköld, Josephsson, & Eliasson, 2004). Persons with a unilateral impairment (e.g., from hemiplegic cerebral palsy and brachial plexus palsy) have a specific condition, with one well-functioning hand and one hand with different degrees of dysfunction. Something as ordinary as playing with building blocks, dressing a doll, zipping up a jacket, or making a sandwich can be quite difficult to manage one-handed. Probably the most important aspect of hand function for persons with unilateral impairments is how effectively the affected hand is used in the performance of activities in which both hands need to be involved. In hand function assessments used in pediatric occupational therapy services, however, commonly one hand is assessed separately from the other, such as in the Melbourne Assessment of Unilateral Upper Limb Function (Randall, Carlin, Chondros, & Reddihough, 2001), Quality of Upper Extremity Skills Test (QUEST) (DeMatteo et al., 1992), and Jebsen–Taylor Hand Function Test (Taylor, Sand, & Jebsen, 1973). Although some hand function assessments involve the use of both hands (e.g., to cut with scissors or to string beads), it is the time or accuracy of the task that is evaluated, not how the tasks are completed concerning the use and role of an affected hand.

The Assisting Hand Assessment

The Assisting Hand Assessment (AHA) is a newly developed test with the purpose of measuring and describing how effectively children who have a unilateral disability use their affected hand (assisting hand) in collaboration with the noninvolved hand during bimanual performance. The AHA was shown to produce valid
measures of assisting hand function for children ages 18 months to 5 years with hemiplegic cerebral palsy or brachial plexus palsy (Krumlinde-Sundholm & Eliasson, 2003; Krumlinde-Sundholm, Holmefur, & Eliasson, 2005). The assessment is based on observations of actions performed in bimanual activity and reflects the person’s habitual performance rather than best capacity.

The AHA is administered in two steps. First, a semistructured 10–15 minute videotaped play session is conducted with toys from the AHA test kit that require bimanual handling. The therapist presents the toys in a playful manner and in a way that ensures that all of the test items can be observed. While playing, the child will display a variety of object-related hand actions. The second part of the administration involves the scoring, which is done from the video recording. The effectiveness of the performance is scored on 22 items using a 4-point rating scale on which 4 = effective, 3 = somewhat effective, 2 = ineffective, and 1 = does not do. The items describe object-related actions of bimanual performance and actions with the assisting hand and are divided into six groups of items: general use, arm use, grasp–release, fine motor adjustments, coordination, and pace. These different aspects of bimanual skills form the concept of how functionally and effectively the affected hand is involved in performing bimanual tasks.

A raw score sum ranging from 22 to 88 is rendered. Equal interval measures in the unit logits are obtained from computing data in a Rasch measurement analysis, which is not yet available for common use. The AHA was developed in Swedish and English versions in parallel. Since its publication in 2003 (Krumlinde-Sundholm & Eliasson, 2003), some minor adjustments have been made to the scale.

Validity and Reliability

The Rasch Measurement Model was used for evaluating evidence of validity and some aspects of reliability of the measures involving 60 assessments (Krumlinde-Sundholm & Eliasson, 2003). Another analysis involving 185 assessments was conducted subsequently (Krumlinde-Sundholm et al., 2005). In both analyses, 95% of the items were found to contribute meaningfully to a unidimensional construct.

Although validity can be seen as the most important aspect of a test, the reliable repeatability of scores between and within assessors also has implications for the validity of score interpretations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Rater reliability is most commonly reported with a reliability coefficient, which is an expression of the extent to which raters can distinguish between individuals on different levels of the measured trait. The reliability coefficient does not indicate, however, how to estimate measurement error for the score of an individual. The standard error of measurement (SEM) is more useful for estimating the score of an individual because it has the same unit of measurement as the actual scores in the test. The SEM can be used to calculate a confidence interval around an observed score within which the true score would lie with 95% probability (American Educational Research Association et al., 1999; Thorndike, 1987). The aim of this study is to evaluate interrater and intrarater reliability for the AHA.

Method

Design

Two different trials of interrater reliability were conducted:

1. In a two-rater design, both raters assessed videotaped AHA sessions with 18 children, using the English version of AHA.

2. In a 20-rater design, all raters rated the same videotaped AHA sessions with 8 children.

In this design, both the English and the Swedish versions of AHA were evaluated, involving 10 English-speaking raters and 10 Swedish-speaking raters. The scores of the two language groups were first analyzed separately. Because no differences were found, these scores were analyzed and reported together and, thus, are referred to as a 20-rater design. Sample sizes of raters and subjects in both interrater designs were based on the guidelines in Walter, Eliasziw, and Donner (1998). A trial of intrarater reliability was conducted of both language versions. Ten English-speaking and 10 Swedish-speaking raters each assessed one videotaped AHA session twice with a 3-week interval. In total, AHA sessions with 8 children were rated for the intrarater trial. The study was approved as a quality assurance project by the Ethics Research Committee at the Karolinska Hospital in Stockholm, Sweden.

Participants

The videotaped AHA play sessions used in this study were selected from a collection of 414 sessions with children between ages 18 months and 71 months. For the two-rater trial, 18 videotaped play sessions were randomly selected. In these sessions, there were 16 children with hemiplegic cerebral palsy and 2 with brachial plexus palsy. The children’s ages ranged from 18 months to 5 years and 11 months. For the 20-rater trial of interrater reliability, 8 children were selected by a stratified random selection, with strata based on function level. The stratification was done to reflect the clinical sample of children for whom this test is intended. In the selected sessions, there were 6 children with hemiplegic
cerebral palsy; with different degrees of impaired hand function, and 2 children with brachial plexus palsy, one with C5-6 lesion and one with C5-Th1 lesion. The children's ages ranged from 19 months to 4 years and 2 months. For the intrarater trial, one of the 8 sessions was randomly assigned to each of the raters; therefore, each session was repeatedly assessed by a minimum of two raters. Informed written consent to use the videotaped AHA sessions for this trial was obtained from the children's parents.

Raters
All raters in this study were occupational therapists working in, or with experience from, pediatric services and were certified AHA raters. Certification requires participation in an instructional 3-day course with subsequent videotaped ratings of 8 children with approved results. Certification courses are given worldwide by the company Handfast AB (www.ahanetwork.se). For the two-rater trial of interrater reliability, two experienced raters were selected. They had each rated more than 30 AHA play sessions before the study. For the 20-rater trial, raters were asked to participate in consecutive order of AHA certification. One requirement for calculating a reliability coefficient is that raters are randomly chosen. Because the AHA is a newly developed test, few raters were available who could be subject to randomization. We therefore decided to invite certified raters to participate in this study in order of their certification and to use the first 10 for each language who agreed to participate. After certification these raters had rated 0–30+ AHA play sessions. Eight raters had rated 0–10 sessions, seven raters had rated 11–25 sessions, and three raters had rated more than 30 sessions before the trial. For the intrarater reliability trial, the same raters as in the 20-rater interrater trial participated.

Statistical Analyses
To compare the scores from the English and Swedish raters in the 20-rater intrarater design, the sum scores were calculated for each child. The median of the sum scores for each language was analyzed with the Wilcoxon matched pairs test. A significance level of 0.01 was used. To compare agreement between the English and Swedish raters in the 20-rater interrater design, the total proportion of agreement for each child within each group of raters was calculated. The total proportion scores for each item in the test then were analyzed with the Wilcoxon matched pairs test (Siegel & Castellan, 1988).

To obtain reliability coefficients, the intraclass correlation coefficient (ICC) was calculated. ICCs were calculated for sum scores and item scores. Raw scores from the AHA are ordinal level data, for which the weighted Kappa would be the most suitable analysis of rater agreement for obtaining reliability coefficients (Altman, 1991; Streiner & Norman, 2003). Because the weighted Kappa with quadratic weights gives equivalent results to the ICC (Fleiss & Cohen 1973), the latter analysis was chosen because it allows for analysis of sum scores from multiple raters. For analysis of interrater reliability, a two-way random effects model was used and referred to as ICC 2,1. According to Shrout and Fleiss (1979), this model is suitable when all subjects are assessed by the same group of raters, and the raters are a sample from a larger group of raters. An absolute agreement definition was used, meaning that the actual scores were analyzed in contrast to the consistency definition that compares the rank ordering given to subjects by raters (Streiner & Norman, 2003). For the intrarater analysis, ICC 1,1 was calculated, and thus a one-way random effects model was used (Shrout & Fleiss, 1979). Confidence intervals were set at 95% in all calculations. The SEM was derived from an analysis of variance analysis (ANOVA) and is the square root of the pooled mean square of MSTime and MSChild*Time. The intrarater SEM was calculated using all ratings in the 20-rater interrater trial, and the intrarater SEM was calculated from repeated ratings of 8 children. The ANOVA analysis for intrarater data also was used to find out whether there was a systematic difference between the first and second rating by each rater.

To analyze whether the experience of different raters in the 20-rater interrater trial was correlated to how they rated the children, several calculations were made. First, a median sum score between raters for each child and the deviation from this median for each rater were calculated. Further, a “gold standard” sum score for each child was obtained by letting AHA course instructors rate the children. From these assessments, median sum scores for each child were calculated, and from these scores each rater’s deviation was calculated. Spearman’s rank order correlation was used to correlate the deviations for each rater with the number of assessments done before the study and with the time interval between taking the instructional course and trial.

Results
The comparison of actual scores between the English and Swedish groups of raters in the 20-rater interrater trial showed that there was no significant difference between the groups for the median of the sum score for each child. The comparison of agreement between the English and Swedish groups of raters in the 20-rater interrater trial showed that there was no significant difference between the groups for any item in the test. Thus, the results from the 20-rater trial will be presented as one trial.
In the 2-rater trial, the interrater ICC for sum score was 0.98 (0.95–0.99) and in the 20-rater trial, 0.97 (0.92–0.99). For individual items, the ICC ranged between 0.35 and 1.00 in the 2-rater trial, with 15 of the 22 items having an ICC higher than 0.70. In the 20-rater trial, the ICC for items ranged between 0.70 and 0.94, with 18 items having an ICC of 0.80 or higher (Table 1). The intrarater ICC for sum score was 0.99 (0.98–0.99). For individual items, the ICC ranged between 0.84 and 1.00, and 15 items had an ICC higher than 0.90 (Table 1).

The SEM for interrater data from the 20-rater design was found to be 1.5, rendering a 95% confidence interval of ± 3 raw scores. The SEM for intrarater data was 1.2, giving a confidence interval of ± 2.4 raw scores. For the intrarater trial there was no systematic change in the way raters scored the children the first and second times.

There were weak correlations between raters’ previous experience with AHA ratings and their deviation from group median ($r = -0.42$), as well as for their deviation from gold standard ($r = -0.34$). There was little or no relationship between time since course completion and their deviation from group median or gold standard ($r = -0.24$). None of these correlations were found to be statistically significant.

**Discussion**

The interrater and intrarater ICCs for sum score are high in all designs, which is a strong indication that results from AHA assessments are reliable. The high ICC also suggests that raters can separate children of different abilities well.

Because the English and Swedish versions of AHA have been developed in parallel, with constant cooperation with native English speakers, both versions are considered to be originals. It was natural, therefore, to conduct the reliability trial in both languages simultaneously. One aim with this particular part of the trial was to compare the interrater reliability between the two language versions. Because there was no significant difference between the two groups of raters, the two versions are considered equivalent.

There is no unified way to interpret reliability coefficients. According to Landis and Koch’s (1977) benchmarks on how to interpret Kappa coefficients, a coefficient of 0.41–0.60 is “moderate,” 0.61–0.80 is “substantial,” and 0.81–1.00 is “almost perfect.” It should, however, be kept in mind that the Kappa coefficient gives lower coefficients than the weighted Kappa, which gives results comparable with the ICC (Fleiss & Cohen, 1973). Polit and Hungler (1999) have provided a general guideline that a reliability coefficient of at least 0.70 (or even better, 0.80) is sufficient for group level comparisons. For making decisions for individuals, a coefficient of 0.90 or more is required. In light of these guidelines, the reliability coefficients found in this study show that the AHA has excellent rater reliability.

The interrater ICC for certain items demonstrated somewhat different results in the 2-rater and 20-rater trials, including the items “chooses assisting hand,” “moves forearm,” and “manipulates.” More than the obvious reason that agreement between raters is different, the different sampling strategies in respective designs highly influenced these results. In the 20-rater design, children were stratified to different ability-levels before sampling; thus, the whole scale is used for all items. In the 2-rater design, children were randomly chosen and by chance the sample did not represent the same variation in ability levels. As a result, the most able children were not represented and the whole scale was not used for the more difficult items. All three items mentioned previously are among the most difficult in the AHA hierarchy (Krumlinde-Sundholm et al., 2005). The way the scale is used has a great effect on the ICC.

### Table 1. Interrater and Intrarater ICC for Sum Score and Items on the Assisting Hand Assessment

<table>
<thead>
<tr>
<th>Item</th>
<th>Interrater Reliability</th>
<th>Intrarater Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Raters</td>
<td>20 Raters</td>
</tr>
<tr>
<td>General use items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approaches objects</td>
<td>1.00</td>
<td>.91</td>
</tr>
<tr>
<td>Initiates use</td>
<td>.77</td>
<td>.89</td>
</tr>
<tr>
<td>Chooses assisting hand</td>
<td>.41</td>
<td>.81</td>
</tr>
<tr>
<td>Arm use items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stabilizes by weight or support</td>
<td>.93</td>
<td>.85</td>
</tr>
<tr>
<td>Reaches</td>
<td>.78</td>
<td>.86</td>
</tr>
<tr>
<td>Moves upper arm</td>
<td>.53</td>
<td>.70</td>
</tr>
<tr>
<td>Moves forearm</td>
<td>.35</td>
<td>.79</td>
</tr>
<tr>
<td>Grasp–release items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grasps</td>
<td>.79</td>
<td>.90</td>
</tr>
<tr>
<td>Holds</td>
<td>.94</td>
<td>.94</td>
</tr>
<tr>
<td>Stabilizes by grip</td>
<td>.86</td>
<td>.85</td>
</tr>
<tr>
<td>Readjusts grip</td>
<td>.77</td>
<td>.88</td>
</tr>
<tr>
<td>Varies type of grasp</td>
<td>.93</td>
<td>.89</td>
</tr>
<tr>
<td>Releases</td>
<td>.75</td>
<td>.94</td>
</tr>
<tr>
<td>Puts down</td>
<td>NAa</td>
<td>.82</td>
</tr>
<tr>
<td>Fine motor adjustment items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moves fingers</td>
<td>.73</td>
<td>.77</td>
</tr>
<tr>
<td>Calibrates</td>
<td>.91</td>
<td>.84</td>
</tr>
<tr>
<td>Manipulates</td>
<td>.46</td>
<td>.80</td>
</tr>
<tr>
<td>Coordination items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordinates</td>
<td>.74</td>
<td>.87</td>
</tr>
<tr>
<td>Orientes objects</td>
<td>.69</td>
<td>.86</td>
</tr>
<tr>
<td>Pace items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proceeds</td>
<td>.67</td>
<td>.77</td>
</tr>
<tr>
<td>Changes strategy</td>
<td>.76</td>
<td>.92</td>
</tr>
<tr>
<td>Flow in bimanual task</td>
<td>.73</td>
<td>.84</td>
</tr>
</tbody>
</table>

*Note: ICC = intraclass correlation coefficient (95% confidence interval)*

*a ICC could not be calculated because all participants received a score of 1 from both raters, and thus no variance was present.*
the whole scale is used, the ICC is higher; if there is less variation in scores given, the ICC is lower. The extreme example of this is the most difficult item, “puts down,” in which all children in the 2-rater design received a score of 1 from both therapists, making it impossible to calculate an ICC.

It is somewhat surprising that the raters’ experience only to a minor degree influenced how they scored the children. The conclusion to be drawn from this material is that the instructional course given to all AHA raters is sufficient to enable them to use the test in a way that ensures interrater and intrarater reliability. The higher ICC from the intrarater trial indicates, however, that when using the test in research studies the same rater should rate all sessions to minimize measurement error.

The result of the SEM and its interval around an observed score is an indication of the variation due to rater. The intrarater interval is to be thought of as the smallest interval within which the true score of the individual lies, and the size of the interval is to be interpreted as the measurement error due to the rater for the score of one AHA play session. The intrarater interval is to be considered if another rater rates the same play session of the same child. The SEM interval of ±2.4 and ±3 raw scores for the intrarater and interrater conditions, respectively, should be seen as acceptable. These numbers are easier to apply to individual results than the ICC information and are important for score interpretation in clinical practice.

Future research to develop AHA would be to study test–retest reliability, which would involve two subsequent sessions with the same child. Validity of AHA measures concerning children in an extended age range of 18 months to 12 years has recently been evaluated (Krumlinde-Sundholm, Holmefur, Kottorp, & Eliasson, in press), and the usefulness of AHA for other diagnostic groups with a unilateral hand dysfunction needs to be investigated. Furthermore, other age groups may be considered; however, the activity context used for observation must be appropriate and meaningful for the respective age in question.

**Conclusions**

This investigation of rater reliability of the AHA shows excellent interrater and intrarater reliability for sum score in all study designs. When the AHA is used for research purposes, it is preferable that the same rater be used because intrarater reliability is higher than interrater reliability. When using the AHA in clinical practice, an interval of ±3 (intrarater) and ±2.4 (intrarater) raw scores should be considered the interval within which the true score lies.

**Acknowledgments**

We wish to express our gratitude to the children and families and to the occupational therapists who participated in this study. The study was financially supported by the Health Care Sciences Postgraduate School, Swedish Research Council, Sällskapet Barnavård, David och Astrid Hagelén Foundation, Norrbacka-Eugenia Foundation, and Frimurare Barnhuset Stockholm Foundation. Parts of this study were presented at the 2005 Annual Meeting of the American Academy for Cerebral Palsy and Developmental Medicine.

**References**


---

**Best-Seller!**

**AOTA’s Self-Paced Clinical Course**

**The Hand**

An Interactive Study for Therapists

*By Judy Colditz, OTR/L, CHT, FAOTA*

This SPCC combines written coursework with interactive computer-based learning and examines the anatomical basis and clinical presentation of problems in the hand and forearm. Using the CD-ROM, *The Interactive Hand: Therapy Edition*, you will explore the multiple layers of complex anatomy while learning about palpation, examination, and the resource of common hand disorders.

This course also is an excellent preparation tool for the Hand Therapy Certification Exam.

Earn 1.6 AOTA CEUs (16 NBCOT PDUs / 16 contact hours).

Order #3017-J (with CD-ROM)

AOTA members: $260  Nonmembers: $360

Call 877-404-AOTA  •  shop www.aota.org (Books, Products, & CE)