Interrater Reliability of the Melbourne Assessment of Unilateral Upper Limb Function for Children With Hemiplegic Cerebral Palsy

Michelle Spirtos, Paul O’Mahony, Jeni Malone

OBJECTIVE. We examined the interrater reliability of the Melbourne Assessment of Unilateral Upper Limb Function.

METHOD. Three occupational therapists independently scored 34 videotaped assessments of children with hemiplegic cerebral palsy aged 6 yr, 1 mo, to 14 yr, 5 mo. Intraclass correlation coefficients (ICCs) at a 95% confidence interval were calculated for total scores, category scores, and item scores.

RESULTS. The correlation between raters’ total scores was high (ICC = .961). The highest correlation for test components between raters was found for fluency (ICC = .902), followed by range of movement (ICC = .866), and the lowest correlation was found for quality of movement (ICC = .683). The ICCs for individual test item scores varied and ranged from .368 to .899.

CONCLUSION. This study demonstrated high interrater reliability for total scores, with scoring of some individual components and items requiring further consideration from both a clinical and a research perspective.


Cerebral palsy (CP) is a clinical term describing a spectrum of neuromuscular disorders. Cases of hemiplegic CP (involvement of one side of the body) constitute the second largest percentage of reported cases (35%; Hagberg & Hagberg, 2000). Children in this group generally present with abnormal tone and decreased quality of movement, speed, and dexterity during functional tasks; reach, grasp, release, and manipulation are impaired in the affected upper limb. Common daily tasks, which typically require the child to use both hands together, can be impaired, including tying shoe laces, manipulating construction toys, and stabilizing paper while writing. The Manual Ability Classification System (MACS; Eliasson et al., 2006) was developed to classify the ability of children with neurological impairment to handle objects in daily activities. It has five classification levels ranging from mild limitation to severe limitation during even simple actions. Because children with hemiplegia have involvement of only one upper limb, they most frequently are categorized as Level 1 (limitations in ease of performance) or Level 2 (decreased quality of performance and speed).

Children with hemiplegic CP are often seen by multiple health care professionals and receive a wide spectrum of interventions, including those with a specific focus on function and quality of movement in the upper limb. Occupational therapy for these children aims to increase quality of movement in the upper limb and to facilitate the child’s acquisition of the functional skills required to participate in the occupations of childhood. Boyd, Morris, and Graham (2001) completed a systematic review of the management of the upper
limb and found the evidence to support the clinical usefulness of most treatment approaches to be limited. The lack of sensitive, reliable, and valid outcome measures is cited as a main contributing factor (Krumlinde-Sundholm & Eliasson, 2003; Siebes, Wijnroks, & Vermeer, 2002). Given that intervention is time consuming and costly to the family and service providers, high-level research and sound measurement tools are essential. Thus, further examination of the available assessments is required to ensure that the results of intervention can be accurately measured and researched by members of the profession.

The Melbourne Assessment of Unilateral Upper Limb Function (MAUULF; Randall, Johnson, & Reddihough, 1999) is a criterion-referenced assessment that was specifically developed to measure unilateral quality of upper-limb movement in children with a neurological impairment (Bourke-Taylor, 2007). It is widely used to examine the effectiveness of specific interventions. The test items have been chosen for their suitability for children and teenagers, and the age range for the assessment is 5–15 yr.

The MAUULF can be administered to children with unilateral or bilateral upper-limb impairment but measures only one hand at a time. It contains 16 items that examine the child’s performance on tasks such as reaching, grasping, pointing, manipulating, releasing, and bringing the hand to the mouth. Individual items are scored under four main categories: range of movement, target accuracy, fluency, and quality of movement. Each item is scored on a scale of either 0–3 or 0–4; the manual provides a detailed description of what is required for each score. Although the assessment manual provides specific scoring criteria, it does not specify the degree of training required by the therapist using the assessment. An independent study by Cusick, Vasquez, Knowles, and Wallen (2005) examined the effect of rater training on the assessment’s reliability and concluded that provision of training to novice users increased familiarity with the test and improved consistency of rating.

The development of the MAUULF was described in an initial reliability study published by Johnson et al. (1994). This study reported a total score interrater reliability of .68. Randall, Carlin, Chondros, and Reddihough (2001) examined the MAUULF’s reliability in a study involving 16 therapists and 20 children in Australia. The authors of this study included the test developers. They concluded that the MAUULF has moderate to high interrater reliability and is useful for comparison between children and for detecting change in hand function after intervention. The total internal consistency of the items was found to be high ($\alpha = .96$). Individual items were also examined for internal consistency, with correlations varying from .90 for Item 15 to .63 for Item 14 and .62 for Item 16.

Examinations of interrater reliability (Randall et al., 1999, 2001) included children with a diagnosis of CP including spastic quadriplegia, diplegia, and ataxia, but they did not include children with hemiplegia. The researchers suggested that the high correlation found depended in part on the heterogeneous sample and hypothesized that reliability would be decreased in a more homogeneous group. Klingels et al. (2008) compared the MAUULF to the Quality of Upper Extremity Skills Test and reported on the MAUULF’s interrater reliability with two raters of 22 children with hemiplegic CP. Details regarding the therapists’ (physiotherapists’) experience and training in the use of the assessment were not provided. Although the MAUULF manual refers to an early sensitivity study, we found no published studies examining the test’s sensitivity or responsiveness to change.

Method

Research Design

In this correlation study, we examined the interrater reliability among 3 raters of the MAUULF. This study took place with ethical approval from the Faculty of Health Sciences ethics committee, Trinity College, Dublin. Written informed consent was obtained from the parents of the children for blind scoring of their child’s assessments as part of a larger study.

Participants

To score the assessments, we recruited a convenience sample of 3 therapists with honors degrees in occupational therapy who were working in the same center for children and adults with physical disabilities. Rater A had 9 yr experience in pediatrics and specialized in children with upper-limb impairments. Both Raters B and C were involved in the upper-limb service provided by the center. Rater B had 7 yr experience (5 in pediatrics), and Rater C had 6 yr experience (5 in pediatrics). All 3 therapists had used the MAUULF as part of their typical practice in the month before the study.

The first author (Spiros) collected the video assessments of the MAUULF as part of a larger study, and she was not involved in the scoring of the assessments. Thirty-four complete video assessments of 11 children were included in the analysis. No data were missing. All children attended services for youth with a physical disability and were recruited as part of a larger intervention study. The
children’s mean age was 8 yr, 4 mo (age range = 6 yr, 1 mo–14 yr, 5 mo). Fifty-five percent of the children (n = 6) were boys and 45% (n = 5) were girls. Through a review of the medical charts, a diagnosis of hemiplegic CP was confirmed for each child. Sixty-four percent (n = 7) had a right hemiplegia and 36% (n = 4) had a left hemiplegia. Fifty-five percent (n = 6) had a MACS level of 1, and 45% (n = 5) had a MACS level of 2.

Procedures
To ensure therapist familiarity with scoring the assessments, the 3 therapist participants completed a 3-hr training session facilitated by Rater A. The therapists initially reviewed the manual together and discussed the scoring criteria. After this discussion, the video of a child (data not included in the study) was played, and all 3 therapists independently scored the assessment. They were permitted to stop the tape and rewatch it at any point but were not allowed to discuss their scores. The group then reviewed and compared the scores, and a consensus was reached regarding the score allocated to each item. They were instructed to review the videotape during the consensus process. They were then given a second video to score, which was reviewed and completed in the same manner (data again not included in the study). All 3 raters were in total agreement on the final scores allocated to each item. The training terminated with a discussion of the assessment, and all 3 therapists reported that they had sufficient knowledge of the assessment to proceed.

The previous study by Randall et al. (2001) reported that a 90-min instruction session was completed by raters, but details were not provided regarding the format or content. Cusick et al. (2005) reported that their trained group received a 2-hr script session, followed by instruction and scoring of each of the 16 items using an instructor and a video of a child.

Each of the 34 anonymous assessments was assigned a case number, and the videos were put on CD. The 3 raters were provided with a pack of CDs and blank MAUULF forms and were requested to rate the assessments individually without further discussion with the other raters.

Data Analysis
We entered the data from the 34 assessments into SPSS Version 14.0 (SPSS Inc., Chicago) and used intraclass correlation coefficients (ICCs) to analyze the assessment’s interrater reliability for each rater’s individual scores. We examined total and item scores and analyzed scores with regard to the four main categories of scoring within the assessment: range of movement, quality of movement, accuracy, and fluency. Because the MAUULF is a criterion-referenced assessment primarily designed to evaluate change after intervention, the manual provides no normative data. Total raw scores are converted to percentages. For the purposes of this study, we used raw scores; the maximum raw score was 122, and the minimum was 0.

We calculated a two-way random effects model using single measures at a 95% confidence interval and absolute agreement (ICC 2, 1 model; McGraw & Wong, 1996; Shrout & Fleiss, 1979). When raters are selected from a probable sample of raters (in this case, pediatric occupational therapists), this model is appropriate because it allows generalizations to be made (Shrout & Fleiss, 1979). The ICC 2, 1 model is a measurement of agreement that provides more useful data than consistency of rating in interrater reliability studies (Eliasziw, Young, Woodbury, & Fryday-Field, 1994; Shrout & Fleiss, 1979). We chose single measures for reporting because they provided more stringent results than do average measures. The choice of the absolute agreement approach over the consistency approach means that all differences in measurement were considered, no matter what the reason for the difference (McGraw & Wong, 1996). This approach provides a more rigorous analysis in this situation than does consistency rating because with the approach adopted (or in the absolute agreement approach), any disagreement between raters was considered in the calculation, whereas consistency rating excludes situations such as one rater being consistently higher than the others.

Because the ICC calculates the agreement between a group of raters, the higher the ICC is, the greater the agreement and the lower the measurement error are (Peat, 2002). Fowler, Jarvis, and Chevannes (2002) stated that values of .40–.69 represent a modest correlation; .70–.89, a strong correlation; and .9–1.00, a very strong correlation. Morris, Kurinczuk, Fitzpatrick, and Rosenbaum (2006) considered ICCs >.70 reliable for sample-based research and required ICCs >.90 for clinical reliability. Neither study made distinctions regarding any variation relating to the method used to calculate the correlation or the ICC.

Results
The ICC for total scores was high (.96), a result indicating a very strong interrater reliability for the test when overall results were considered. Of the four category total scores, range of movement, target accuracy, and fluency had ICCs >.70; fluency demonstrated a very strong correlation (.90). Quality of movement had a correlation of .68. When each item was individually analyzed, only 9 of the 16 items achieved an ICC >.70. Item 14 (hand-to-hand
transfer) received the lowest ICC (.37). The ICC scores and confidence interval results are listed in Table 1.

**Discussion**

Using the Fowler et al. (2002) guidelines, the ICCs for the 3 raters’ total scores represent a strong correlation, indicating that the test results of the MAUULF are clinically reliable. Our results compare favorably with those of previous studies (Cusick et al., 2005; Randall et al., 2001) and indicate that with a homogeneous group (hemiplegic CP), the interrater reliability remains high.

Of interest is that the interrater reliability results for the assessment components demonstrated a strong correlation for range of movement and target accuracy and a very strong correlation for fluency. Only the result for quality of movement fell below .70. An examination of the specific components of the test has not previously been published. Scoring in quality of movement may be more subjective than scoring in the other components. The test items that measure performance specifically with regard to quality of movement relate to precision of release. On review of the literature, many published studies relate to object grasp, whereas a relative paucity exists with regard to object release, an important component of hand function (Eliasson & Gordon, 2000). The discrepancy likely reflects the focus of current practice. Thus, the level of expertise in assessing performance during object release may need to be addressed.

Correlation results for individual items were lower than those obtained in previous studies: Only 9 of the 16 items received an ICC > .70. The lower band of the 95% confidence interval drops below .70 for 12 of the 16 items. Item 14, hand-to-hand transfer, achieved the lowest correlation result (.37); note, however, that scoring for this item was from only one contributing score, whereas most items have three or more contributing scores. This item received the second lowest item ICC in the Randall et al. (2001) study. The test manual also reported a low correlation for Item 14, which relates to hand-to-hand transfer and is scored on a scale ranging from 0 to 4 (Randall et al., 1999). The differences between the individual scoring criteria for each point are small, however, and further clarity within the test manual may be needed to increase agreement between raters. The test manual argues for the continued inclusion of this item, but in light of the findings of recent studies, it may be worth reconsidering its inclusion.

The Cusick et al. (2005) study was designed to examine the effect of training on rating of the MAUULF by novice raters (students due to graduate in 12 wk). Agreement between raters was higher than in our study, where the ICC was <.70 for 12 of the 16 items. The reason for this significant difference is not clear; however,

<table>
<thead>
<tr>
<th>Melbourne Assessment of Unilateral Upper Limb Function Components</th>
<th>ICC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>.96</td>
<td>.93–.98</td>
</tr>
<tr>
<td>Total range of movement</td>
<td>.87</td>
<td>.77–.93</td>
</tr>
<tr>
<td>Total quality of movement</td>
<td>.68</td>
<td>.52–.81</td>
</tr>
<tr>
<td>Total target accuracy</td>
<td>.77</td>
<td>.66–.88</td>
</tr>
<tr>
<td>Total fluency</td>
<td>.90</td>
<td>.83–.95</td>
</tr>
<tr>
<td>Item 1: Reach forward (range of movement, target accuracy, fluency)</td>
<td>.68</td>
<td>.52–.81</td>
</tr>
<tr>
<td>Item 2: Reach forward to an elevated position (range of movement, target accuracy, fluency)</td>
<td>.62</td>
<td>.43–.77</td>
</tr>
<tr>
<td>Item 3: Reach sideways to an elevated position (range of movement, target accuracy, fluency)</td>
<td>.62</td>
<td>.43–.77</td>
</tr>
<tr>
<td>Item 4: Grasp crayon</td>
<td>.77</td>
<td>.64–.87</td>
</tr>
<tr>
<td>Item 5: Drawing grasp</td>
<td>.61</td>
<td>.40–.77</td>
</tr>
<tr>
<td>Item 6: Release of crayon (range of movement, quality of movement, accuracy of release)</td>
<td>.90</td>
<td>.83–.94</td>
</tr>
<tr>
<td>Item 7: Grasp of pellet</td>
<td>.87</td>
<td>.80–.93</td>
</tr>
<tr>
<td>Item 8: Release pellet</td>
<td>.65</td>
<td>.48–.80</td>
</tr>
<tr>
<td>Item 9: Manipulation (finger dexterity, fluency)</td>
<td>.77</td>
<td>.64–.87</td>
</tr>
<tr>
<td>Item 10: Pointing (red, green, yellow, blue square)</td>
<td>.80</td>
<td>.68–.89</td>
</tr>
<tr>
<td>Item 11: Reach to brush from forehead to back of neck (range of movement, fluency)</td>
<td>.76</td>
<td>.62–.86</td>
</tr>
<tr>
<td>Item 12: Palm to bottom (range of movement, fluency)</td>
<td>.90</td>
<td>.83–.95</td>
</tr>
<tr>
<td>Item 13: Pronation/supination</td>
<td>.86</td>
<td>.77–.92</td>
</tr>
<tr>
<td>Item 14: Hand-to-hand transfer</td>
<td>.37</td>
<td>.16–.58</td>
</tr>
<tr>
<td>Item 15: Reach opposite shoulder (range of movement, target accuracy, fluency)</td>
<td>.78</td>
<td>.63–.88</td>
</tr>
<tr>
<td>Item 16: Hand-to-mouth and down (range of movement, target accuracy, fluency, speed)</td>
<td>.70</td>
<td>.40–.84</td>
</tr>
</tbody>
</table>
we can make several comments. When the populations of the studies were examined, we noted that more than half of the children videotaped in the Cusick et al. study had a diagnosis of CP with spastic quadriplegia, whereas all the children in our study had a diagnosis of hemiplegic CP. Thus, the children in Cusick et al.’s study had a greater level of impairment and were arguably more easily rated than the children in our study, in which more than half of the children were classified at MACS Level 1, indicating greater skill and more subtle limitations.

Experience working as a pediatric therapist also may have an effect on how a therapist scores a quality-of-movement assessment. Sorsdahl et al. (2008) examined observer reliability on the Gross Motor Performance Measure (Boyce et al., 1991) and the Quality of Upper Extremity Skills Test (DeMatteo et al., 1992). Although both raters had substantial pediatric experience, Sorsdahl et al. (2008) found that the rater with less experience (10 yr) had more reliability coefficients <.70 than did the rater with more experience (17 yr). They suggested that novice raters are not sufficiently qualified to make objective judgments regarding quality of movement. Interestingly, the novice raters in the Cusick et al. (2005) study had higher item ICCs than did the experienced raters in this study. The novice raters may have had limited exposure to a pediatric sample and had not used the assessment before the experiment; they therefore had not developed preconceived methods of viewing the child that persisted even after new training in the assessment took place.

Although we found some small differences between the training provided to the raters in this study and that outlined in the previous studies, they do not appear to be sufficient to warrant the significantly lower results for individual items. In the Sorsdahl et al. (2008) study examining reliability of the Gross Motor Performance Measure and the Quality of Upper Extremity Skills Test, the raters completed 22 hr of training over a 4-mo period, an amount significantly greater than that completed by raters in any study examining the MAUULF. Whether this degree of training is achievable or realistic in a clinic environment is questionable.

This study had several limitations. The number of participants was relatively small, and a larger sample could have led to stronger conclusions. We used a convenience sample of raters and children.

In summary, we found that after a training session in the scoring of the MAUULF, raters’ total scores showed a strong interrater reliability. The quality-of-movement component highlighted difficulties with scoring related to release, and this point is important both clinically and for future research. The test developers should further consider the inclusion of Item 14. This study supports the use of this test as a tool to measure change after intervention and to compare different children’s performance or response to treatment; however, additional studies examining the sensitivity of the test and responsiveness over time are required.

Acknowledgments
We extend a most sincere thank you to all the children who participated in this study. This research was partially funded by a research grant from the Central Remedial Clinic, Dublin, Ireland.

References


