J Rehabil Med 2006; 38: 118–123

Taylor & Francis
Taylor & Francis Group

INTRA- AND INTER-RATER RELIABILITY OF THE ASSESSMENT OF CAPACITY FOR MYOELECTRIC CONTROL

Liselotte M. Hermansson, OT, PhD[1,2], Lennart Bodin, PhD[3] and Ann-Christin Eliasson, OT, PhD[2]

From the [1]Limb Deficiency and Arm Prosthesis Centre, Örebro University Hospital, Örebro, [2]Department of Woman and Child Health, Karolinska Institute, Stockholm and [3]Unit of Statistics and Epidemiology, Clinical Research Centre, Örebro University Hospital, Örebro, Sweden

Objective: **To examine the reliability of the Assessment of Capacity for Myoelectric Control (ACMC) in children and adults with a myoelectric prosthetic hand.**
Design: **Intra-rater and inter-rater reliability estimated from reported assessments by 3 different raters.**
Patients: **A sample of convenience of 26 subjects (11 males, 15 females) with upper limb reduction deficiency or amputation and myoelectric prosthetic hands were video-taped during a regular clinical visit for ACMC. Participants' ages ranged from 2 to 40 years.**
Methods: **After instruction, 3 occupational therapists with no, 10 weeks' and 15 years' clinical experience of myoelectric prosthesis training and follow-up independently rated the 30 ACMC items for each patient. The ratings were repeated after 2–4 weeks. Inter- and intra-rater reliability in items was examined by using weighted kappa statistics and Rasch-measurement analyses.**
Results: **The mean intra-rater agreement in items was excellent (kappa 0.81) in the more experienced raters. Fit statistics showed too much variation in the least experienced rater, who also had only good (kappa 0.65) agreement in items. The stability of rater calibrations between first and second assessment showed that no rater varied beyond chance (>0.50 logit) in severity. The mean inter-rater agreement in items was fair; kappa 0.60, between the experienced raters and kappa 0.47 between raters with no and 10 weeks' experience.**
Conclusion: **Overall, the agreement was higher in the more experienced raters, indicating that reliable measures of the ACMC require clinical experience from myoelectric prosthesis training.**

Key words: reproducibility of results, measurement, arm prosthesis, occupational therapy.

J Rehabil Med 2006; 38: 118–123

Correspondence address: Liselotte Hermansson, Limb Deficiency and Arm Prosthesis Centre, Örebro University Hospital, SE-701 85 Örebro, Sweden. E-mail: liselotte.hermansson@orebroll.se

Submitted January 7, 2005; accepted August 7, 2005

INTRODUCTION

The purpose of this study was to examine the reliability of the Assessment of Capacity for Myoelectric Control (ACMC) (1), a recently developed observation-based assessment that measures a person's capacity to control a myoelectric prosthetic hand during the performance of ordinary daily tasks.

Reliability refers to the consistency of measurements when the procedure is repeated on a population of individuals or groups (2). The need for standardized, observational assessments of the performance of a person with a myoelectric prosthetic hand has been pointed out (3–5). However, the first step to be taken before the prosthesis can be actively used in the performance of daily activities is to learn the ability to control the prosthesis. The capacity to control a myoelectric prosthetic hand is essential for the future use of the hand in daily life (6), and in the light of this fact the ACMC was developed.

The ACMC is a test based on clinical observations of the clients, which can be made when the client is performing any task involving the use of 2 hands. The 30 items comprising the ACMC represent different levels of capacity for control of the myoelectric hand when gripping, holding and releasing daily life objects. An earlier study has demonstrated the hierarchical order of the items, showing how they range from easy to hard (1), making it possible to evaluate clients with varying degrees of ability. By Rasch measurement analysis (7, 8) the data are converted into linear measures, thus combining the person's ability and the item's difficulty in a probabilistic model.

In the previous study mentioned above, it was shown that the ACMC was sensitive enough to evaluate changes over time in groups of clients with myoelectric prosthetic hands. However, when occupational therapists use the ACMC in their work, it is important to determine whether the ACMC can score consistently, both within and between raters. Thus, for further use of the ACMC in clinical practice, the reliability of the instrument needed to be determined.

The present study was therefore undertaken to evaluate this instrument regarding intra- and inter-rater reliability. The specific research questions addressed were as follows: (*i*) Do the raters display consistent scoring in repeated assessments?; (*ii*) Is the scoring consistent between raters?; and (*iii*) Are there any indications of a pronounced difference between inexperienced and more experienced raters?

 ISSN 1650-1977
DOI: 10.1080/16501970500312222

METHODS

Design

To evaluate the reliability of the ACMC, 26 persons were video-taped over a 4-month period. They were filmed during a regular visit to the limb-fitting centre for training in or follow-up of the use of a myoelectric prosthesis. Three independent raters made the assessments on the basis of the persons' performance as seen on the videos. Each subject-video was rated in the same order by the 3 raters. For the intra-rater evaluation, all assessments were repeated in the same order 3–4 weeks later by each of the raters; thus there were 2 sessions of ACMC assessments in this analysis.

The local Ethics Committee approved the study. In addition, oral consent was received from the subjects and, in younger subjects, their parents gave consent.

Subjects

The subjects comprised a sample of convenience of 11 males and 15 females (mean age 10 years; range 2–40 years) with a myoelectrically controlled prosthesis. They were recruited from patients attending the limb-fitting centre during the period August to December 2002. One subject, a female, attended the centre twice with a 3-month interval during the period. Since she was performing different tasks and the situations in which the tasks were performed were different in these occasions the videos were considered to be non-dependent and, hence they were both used. An effort was made to recruit patients with varying degrees of capacity for myoelectric control. The subjects had had the prosthesis for a mean period of 6 years (range 0–20 years).

Raters

To represent new users of the instrument, one randomly assigned occupational therapy student (rater A) with no previous experience of myoelectric prosthetic training and one occupational therapy student with 10 weeks' practice at the limb-fitting centre, i.e. with some experience of myoelectric prosthetic training and ACMC assessments (rater B), both in their last year of education, were trained in the ACMC method. To represent experienced users, one of the most experienced occupational therapists (rater C) at the limb-fitting centre, with previous training in the ACMC method, was assigned to this study. All raters received the same information and they all had a copy of the ACMC manual (9).

Instrumentation

The ACMC is scored on the basis of observations of the myoelectric prosthesis user as he or she is performing everyday tasks. Any task, easy or difficult, can be used to evaluate the capacity for control as long as the task requires active use of both hands (i.e. the unaffected hand and the prosthetic hand). During the assessment, the subjects are encouraged to accomplish the tasks spontaneously in their usual way (i.e. by using the prosthetic hand as they are used to, as an active assisting hand or as a passive support or stabilizer of objects). The occupational therapist assesses their capacity for control of their myoelectric prosthesis by rating their performances on 30 items representing different aspects of quality of myoelectric control. The 30 items in the ACMC are classified into 4 groups: (*i*) gripping (12 items), (*ii*) holding (6 items), (*iii*) releasing (10 items), and (*iv*) co-ordinating between hands (2 items) (1).

Each person's performance is rated with scores ranging from zero to 3, where 0 = not capable, 1 = sometimes capable, capacity not established, 2 = capable on request, and 3 = spontaneously capable. Only those items that are observed during the test session are scored. In accordance with Rasch measurement models, items not observed are recorded as missing; the estimation of item and person statistics when using Rasch models allows for missing data (10). To convert the ordinal ratings into linear measures, Rasch measurement analysis according to a rating scale model with 4 response categories is performed (10).

Rasch measurement analyses is a family of methods based on a probabilistic relation between any items' difficulty and any persons' ability. A central property of Rasch models is the logit (log-odds-probability units) representation of these concepts, which leads to a characteristic unique for Rasch modelling, namely parameter separation (11). The subject's abilities are represented independently of the specific items and item difficulty independent of specific samples. Rasch models for both dichotomous items as well as ordered response categories have been used increasingly in rehabilitation to develop linear measures of ability (1, 12–14).

Data analysis

The data were analysed in 2 ways. First, analysis of individual ACMC items concerning inter- and intra-rater reliability was performed using the weighted kappa statistic with weights according to the quadratic model (15). The ACMC instrument has 4 response categories, thus the weights used for the analyses were 1.0, 0.889, 0.556 and 0.00. The kappa estimates were supplemented with 95% confidence intervals (CI) (15). The upper limit of the CI was truncated at 1.00, in case the standard formula had given values above the theoretical upper limit of kappa. The guidelines for the interpretation of kappa proposed by Fleiss et al. (15) were used to interpret the strength of the agreement. Agreement below kappa 0.40 was considered poor, and it was judged that those items would probably need further definition to enhance the agreement. To summarize the item estimates of kappa we calculated their mean and will refer to this in the following as mean kappa.

Next, each subject's assessments (27 video recordings $\times 3$ raters $\times 2$ sessions $= 162$ assessments) were analysed using the many faceted Rasch analysis (16), applying the computer software FACETS (version 3.49) according to a 3-facet rating scale model (17). The following 3 facets were considered in the analysis: (*i*) the capacity of the persons, (*ii*) the severity of the rater, and (*iii*) the difficulty of the items. The Rasch analyses are reported with estimates of measures for the subject's capacity and calibrations for the rater's severity, supplemented with the standard error of the calibrations. Measures and calibrations are expressed in logits. In Rasch analysis, goodness-of-fit statistics are used to indicate the degree to which each rater's ordering of persons is consistent with the estimated subject ability measures (intra-rater reliability) (8). In this study the criteria for acceptable rater reliability were $0.6 \leq$ mean-square (MnSq) residuals ≤ 1.4 and/or $-2 < z < 2$, the same as were used in the development of the ACMC (1). Another way of analysing intra-rater reliability is to use the rater severity calibrations and look at the stability of calibrations over time. The estimated difference between raters' severity calibration in 2 sessions gives an indication of the intra-rater reliability. This has been done in several studies in larger populations (14, 15) but is not readily applicable to the sample in the present study. However, the comparison may add some valuable information and was therefore carried out nevertheless.

The model was applied in 2 different settings. The first aimed at a global analysis of both intra- and inter-rater reliability. Each of the 2 sessions was analysed separately in order to obtain comparable results. Parameters for each one of the 3 raters and the 2 sessions were estimated and evaluated. By means of fit statistics and comparisons of calibrations from each session, intra-rater reliability was estimated. The second formulation of the Rasch analysis simplified the model to include only one rater, in order to obtain individual measures for each subject for each rater and time (session). This analysis was repeated for all 3 raters. The session-wise differences in the resulting subject measures were plotted against their mean for the 2 sessions to obtain a 'Bland-Altman plot' (20).

RESULTS

Since the ACMC allows for missing items, some items were scored in almost every subject, whereas other items were often left blank. Overall, the less experienced raters (raters A and B) left many more blanks than the experienced rater (rater C; Fig. 1). In addition, during the first session rater A did not realize that 1 of the videotapes (subject #12) was not fully rewound. Hence, she missed some information and scored many blanks in that subject. All 3 raters scored most items more frequently during the second session than during the first, indicating that by repeating the assessments the raters had improved their ability to observe the items.

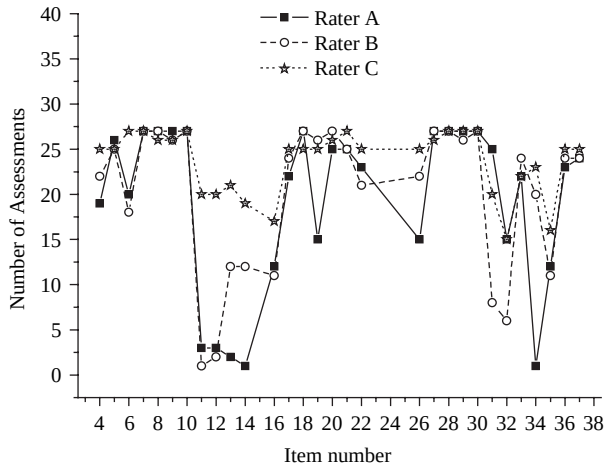

Fig. 1. Number of assessments performed by 3 raters for 30 items of the Assessment of Capacity for Myoelectric Control protocol, session 2 (no items of numbers 1–3, 15, or 23–25). Rater A (inexperienced, rater B (some experience), rater C (experienced).

Intra-rater agreement

Analysis of the individual ACMC items showed that the intra-rater agreement in the more experienced raters (raters B and C; mean kappa 0.81 for both raters) was higher than in the non-experienced rater (rater A; mean kappa 0.65) (Table I). In addition, the rater goodness-of-fit statistics (Table II) indicated that rater A had too many variations in her ratings (MnSq > 1.4). According to the fit statistics for session 1, raters B and C, in contrast, showed consistency in scoring and thus strong intra-rater reliability. In session 2, however, rater B tended to limit her use of the rating scale; MnSq <0.6 indicating fewer variations than expected. The result from the Rasch analysis (Table II), moreover, showed that the severity calibration difference between session 1 and session 2 was smallest for rater B, whereas raters A and C had somewhat larger differences, though of different directions. Although not readily applicable to this data set, the differences in severity calibration between sessions 1 and 2 are well within the limits reported by others (19), indicating stability in rater severity.

Another aspect of intra-rater agreement is illustrated in Fig. 2. Here, the individual measures for each subject are analysed in the Bland-Altman plot. For rater C there is a constant difference, close to zero, between individual measures in the 2 sessions, whereas this is not the case for the less experienced raters. Rater A, the inexperienced rater, scored higher at the second session in the least able persons, and scored lower at the second session for the most able persons. The shift was substantial, as seen from the slope of the regression line. Rater B had a similar performance but in the opposite direction.

The intra-rater agreement in individual items in rater A ranged from kappa 0.32 to 0.95, in rater B from kappa 0.06 to

Table I. *Weighted kappa with 95% confidence interval (within parentheses) of intra-rater agreement for Assessment of Capacity for Myoelectric Control ratings. Duplicate assessments by each rater*

Item category	Item no.	Rater A	Rater B	Rater C
Gripping	4	0.49 (0.12–0.86)	0.89 (0.78–1.00)	0.44 (0.00–0.89)
	5	0.32 (0.07–0.57)	0.94 (0.84–1.00)	0.53 (0.17–0.88)
	6	0.82 (0.66–0.98)	0.06 (−0.40–0.52)	0.77 (0.61–0.94)
	7	0.80 (0.60–1.00)	0.92 (0.82–1.00)	0.76 (0.56–0.96)
	8	0.62 (0.38–0.87)	0.95 (0.89–1.00)	0.90 (0.81–0.99)
	9	0.71 (0.43–1.00)	0.84 (0.75–0.93)	0.85 (0.69–1.00)
	10	0.95 (0.91–1.00)	0.89 (0.80–0.97)	0.84 (0.67–1.00)
	11	n.d.	n.d.	0.81 (0.52–1.00)
	12	n.d.	n.d.	0.74 (0.42–1.00)
	13	n.d.	n.d.	0.80 (0.61–0.99)
	14	n.d.	n.d.	0.80 (0.61–1.00)
	16	0.81 (0.65–0.97)	0.64 (0.20–1.00)	0.87 (0.66–1.00)
Holding	17	0.41 (0.05–0.77)	0.90 (0.79–1.00)	0.80 (0.55–1.00)
	18	0.34 (−0.09–0.77)	0.94 (0.87–1.00)	0.69 (0.37–1.00)
	19	0.58 (0.10–1.00)	0.87 (0.77–0.97)	0.91 (0.84–0.99)
	20	0.68 (0.36–1.00)	0.90 (0.81–0.99)	0.80 (0.63–0.96)
	21	0.63 (0.21–1.00)	0.90 (0.77–1.00)	0.87 (0.74–1.00)
	22	0.66 (0.31–1.00)	0.79 (0.56–1.00)	0.87 (0.71–1.00)
Releasing	26	0.32 (−0.04–0.67)	0.93 (0.83–1.00)	0.77 (0.48–1.00)
	27	0.76 (0.48–1.00)	0.93 (0.85–1.00)	0.84 (0.76–0.93)
	28	0.61 (0.32–0.91)	0.83 (0.74–0.93)	0.93 (0.87–0.98)
	29	0.47 (0.09–0.85)	0.86 (0.75–0.96)	0.95 (0.90–1.00)
	30	0.83 (0.63–1.00)	0.85 (0.75–0.96)	0.93 (0.84–1.00)
	31	0.42 (−0.04–0.89)	0.80 (0.53–1.00)	0.94 (0.87–1.00)
	32	0.88 (0.70–1.00)	0.50 (−0.25–1.00)	0.87 (0.65–1.00)
	33	0.85 (0.64–1.00)	0.87 (0.76–0.98)	0.80 (0.63–0.97)
	34	n.d.	0.91 (0.76–1.00)	0.86 (0.72–1.00)
	35	0.70 (0.40–1.00)	0.64 (0.20–1.00)	0.83 (0.64–1.00)
Co-ordinating	36	0.84 (0.67–1.00)	0.78 (0.60–0.96)	0.79 (0.60–0.97)
	37	0.79 (0.61–0.98)	0.77 (0.59–0.94)	0.79 (0.60–0.97)
Kappa (mean value)		0.65	0.81	0.81

n.d. = items could not be analysed because of the small number of ratings on these items.

Table II. *Rater severity calibration difference for Assessment of Capacity for Myoelectric Control in 2 sessions*

	Session 1 (S1)				Session 2 (S2)				
Rater	Rater everity (logits)	SE (logits)	Infit MnSq	z Std	Rater severity (logits)	SE (logits)	Infit MnSq	z Std	Calibration difference S1 −S2 (logits)
B	0.27	0.06	0.79	−3.2	0.28	0.05	0.45†	−9.0	0.01
A	0.24	0.09	1.62*	7.2	0.08	0.08	2.05*	9.0	−0.16‡
C	−0.51	0.06	0.82	−2.8	−0.37	0.05	0.75	−4.5	0.14

*Rater who assigned unexpectedly high or low scores; †Rater who tended to limit her use of the range of the rating scale; ‡Negative difference in rater severity calibration indicates greater rater severity in session 2. SE: standard error. For further explanation, see Data analysis.

0.95, and in rater C from kappa 0.44 to 0.95 (Table I). In all raters, the lower intra-rater item agreement was noted in the easiest gripping items (items 4, 5 and 6). In raters A and B, the highest intra-rater item agreement was also in gripping items (numbers 8 and 10), whereas in rater C the highest intra-rater item agreement was found in a releasing item (number 29). For items 5, 18 and 26 in rater A, and item number 6 in rater B, intra-rater kappa was $\leq$0.40. In rater C there was no item with intra-rater kappa $\leq$0.40. This indicates that in inexperienced raters items 5, 6, 18 and 26 are more likely to be inconsistently rated.

Inter-rater agreement

Because of the missing information for rater A, session 1, the results from the second session were used for analyses of the inter-rater agreement.

Overall, in individual items the agreement between rater B and rater C (mean kappa 0.60) was higher than that between rater A and rater B (mean kappa 0.44). The agreement in individual items between raters A and B ranged from kappa −0.01 to 0.71, and between raters B and C from kappa 0.04 to 0.84 (Table III). Again, the lowest inter-rater agreement was found for the easier gripping items (items 4 and 6). The highest agreement between both raters A and B and raters B and C was noted for the releasing item number 30. In raters A and B there were 8 items with inter-rater kappa $\leq$0.40 (items 4, 6, 9, 16, 17, 19, 22 and 35). In 3 of these items (6, 16 and 35) the agreement was also low between raters B and C. For all other items the inter-rater kappa value for B and C was >0.40 (Table III).

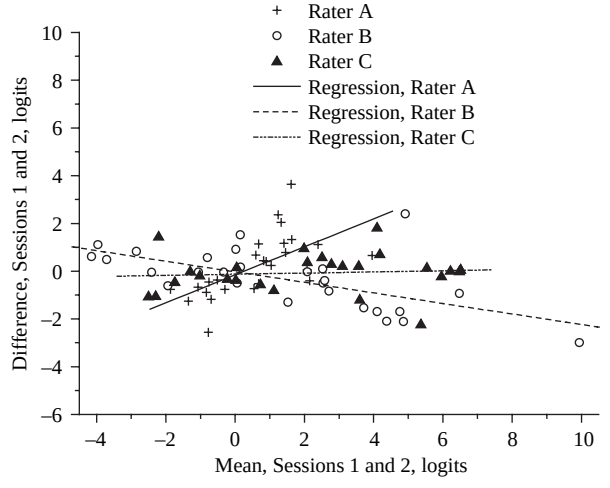

Fig. 2. Intra-rater reliability illustrated by a Bland-Altman plot. Pairwise differences in subject measures (logits) between 2 different sessions are plotted against their means. Linear regression lines are given to show the trend. A horizontal trend located at 0 indicates no systematic difference between sessions. Rater A (inexperienced), rater B (some experience) and rater C (experienced).

The Rasch analysis for inter-rater agreement was based on only 3 raters, for which reason we concentrated on the analysis of the individual measures for each subject, and these are illustrated in the Bland-Altman plot in Fig. 3. The figure shows that rater A differed in a systematic way from both rater B and rater C, since the difference between the raters was dependent on the size of the individual measures. The difference between raters B and C was less systematic and on average close to zero, a preferred result in comparisons of this kind (20).

DISCUSSION

In this study we found an intra-rater reliability that among the experienced raters was on average almost perfect, and in the rater with no clinical experience was substantial. The inter-rater agreement, however, was moderate between both groups of raters. These results indicate that however small, the clinical experience in rater B meant that she was in greater agreement with the more experienced rater than with her fellow student. Also, not surprisingly, it was evident that a substantial training period and clinical experience are necessary for consistent use of the ACMC.

In comparison with other Rasch-derived instruments such as, for example, the Assessment of Motor and Process Skills (13), the methods used for rater reliability analyses in this study (kappa statistics and Rasch analysis) derived from different psychometric traditions. We found both methods very useful, since they added different perspectives to the study. Besides the overall rater agreement, kappa statistics identified certain ACMC items that need further clarification for use by less experienced raters (Tables I and III). In the same way, besides fit statistics from the Rasch analyses, the Bland-Altman plot, demonstrating the stability in subject measures obtained by rater C (Fig. 2) added useful information.

Table III. *Weighted Kappa with 95% confidence interval (within parentheses) of inter-rater agreement for Assessment of Capacity for Myoelectric Control ratings*

Item category	Item no.	Raters A and B	Raters B and C
Gripping	4	0.11 (−0.21–0.44)	0.47 (0.15–0.78)
	5	0.51 (0.27–0.76)	0.53 (0.26–0.81)
	6	0.38 (−0.09–0.84)	0.04 (−0.41–0.50)
	7	0.66 (0.43–0.89)	0.81 (0.65–0.97)
	8	0.63 (0.38–0.88)	0.80 (0.65–0.95)
	9	0.30 (0.02–0.89)	0.73 (0.58–0.88)
	10	0.69 (0.50–0.87)	0.77 (0.60–0.93)
	11	n.d.	n.d.
	12	n.d.	0.67 (0.36–0.98)
	13	0.43 (0.09–0.77)	0.52 (0.21–0.82)
	14	n.d.	0.43 (−0.04–0.89)
	16	0.33 (−0.05–0.71)	0.33 (−0.02–0.68)
Holding	17	0.28 (0.02–0.54)	0.52 (0.21–0.83)
	18	0.54 (0.28–0.79)	0.70 (0.50–0.90)
	19	0.24 (−0.18–0.66)	0.59 (0.37–0.81)
	20	0.42 (0.07–0.76)	0.64 (0.40–0.88)
	21	0.43 (0.18–0.68)	0.75 (0.57–0.93)
	22	−0.01 (−0.39–0.37)	0.73 (0.55–0.91)
Releasing	26	0.53 (0.27–0.78)	0.63 (0.33–0.93)
	27	0.66 (0.46–0.87)	0.67 (0.48–0.85)
	28	0.55 (0.28–0.82)	0.74 (0.57–0.90)
	29	0.43 (0.14–0.73)	0.77 (0.61–0.92)
	30	0.71 (0.54–0.88)	0.84 (0.75–0.93)
	31	0.50 (0.30–0.70)	0.51 (0.07–0.94)
	32	n.d.	0.78 (0.50–1.00)
	33	0.47 (0.17–0.77)	0.46 (0.12–0.80)
	34	n.d.	0.73 (0.50–0.96)
	35	0.33 (−0.05–0.71)	0.37 (−0.02–0.76)
Co-ordinating	36	0.48 (0.20–0.77)	0.44 (0.12–0.75)
	37	0.49 (0.19–0.78)	0.46 (0.15–0.78)
Kappa (mean value)		0.44	0.60

n.d. = items could not be analysed because of the small number of ratings on these items.

The video-recordings used for the analyses had some shortcomings. In ACMC assessments in clinical practice, the patients are observed by the occupational therapist during the performance of different tasks. In this study, the assessments were made on the basis of information from the videos only. This meant that since the information available for the rater was limited to what was in the video, some items might have been difficult to identify. This may explain the lack of scoring on some of the items. For example, the video-operator may have focused on the hands, zoomed in to them, and thus missed the information on how the client was using his/her sight to compensate for the lack of sensation. This is clearly demonstrated by the small number of ratings on items representing use of visual feedback for control of the prosthetic hand (items 13, 14, 16, 34 and 35) (Fig. 1). Furthermore, in the videos during the task performances the easiest gripping items (numbers 4, 5 and 6) were not always shown in the most able persons. In these cases of lack of information it seems as if the raters used different strategies, as demonstrated by the low intra- and inter-rater agreement in these items. Scoring of items 11 and 12 also appeared to be very difficult, especially for the less experienced raters. Revision of the ACMC manual to clarify the importance

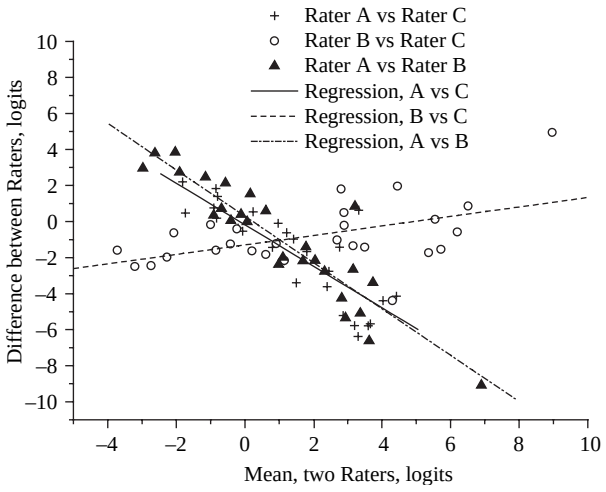

Fig. 3. Inter-rater reliability illustrated by a Bland-Altman plot. Pairwise differences in subject measures (logits) between 2 raters are plotted against their means. Linear regression lines are given to show the trend. A horizontal trend located at 0 indicates no systematic difference between raters. Rater A (inexperienced), rater B (some experience) and rater C (experienced).

of scoring all observable items and then implementing further research is warranted.

The differences between inexperienced and well-trained raters that have been demonstrated in this study clearly illustrate the importance of training and experience for consistent ratings both within and between raters. Firstly, the more experienced the rater, the fewer the items that were inconsistently rated, as seen in Table I. Secondly, the most experienced rater showed no systematic difference between sessions 1 and 2 (Fig. 2), and thirdly, there was a systematic difference both between raters A and B and between raters A and C in subject measures (Fig. 3). Moreover, the less experienced raters had more missing items than the more experienced rater. The results suggest that more than 10 weeks of practice are required for consistent ACMC ratings.

In the future use of the ACMC, only raters with some experience from myoelectric control training will be recommended as users of the instrument. Thus, when considering how to handle the items with low rater reliability in this study, we decided to consider only results from raters B and C. In these raters there were 3 items (items no 6, 16 and 35, Tables I and III with kappa <0.40. Omitting of these items was considered. However, due to the shortcomings in the video-recordings we decided to keep them for future analyses and consideration. In forthcoming studies, the degree of experience for consistent ACMC ratings will also be analysed.

In several studies (18, 19) the variability between raters in terms of severity or leniency has been demonstrated. Raters seem to establish an individual profile of severity and usually tend to maintain this across clients and protocols (18). The results from this study, with rater severity varying, in session 1 from −0.51 (lenient) to 0.27 (severe) logits, and in session 2 from −0.37 to 0.28 logits (Table II), and with a calibration difference of less than 0.16 logits, are in line with those findings.

This variation between raters in their manner of rating, and in their variability in calibration severity, may partly explain the low inter-rater agreement (kappa 0.47 and 0.60) shown in this study. There was no logic, however, in the difference in severity between raters with different degrees of experience (Table II), which indicates that variability in severity is more dependent on rater personality than on experience.

The impact of tasks and rater severity on subject ability measures has been described earlier (8). Besides the judgement of the specific rater, the items may not be equally difficult to perform in different situations or tasks (e.g. feeding, cooking, crafts). These are factors that need to be considered and will require evaluation with a larger sample.

The results from this study have given indications of how much experience and training is needed for reliable measures with the ACMC. A further study on larger populations to address both rater calibration stability and rater severity is in progress.

In conclusion, until the ACMC can adjust for rater severity we recommend that for clinical trials or follow-up, the same rater should perform the ACMC. The assessment method requires training and practice.

ACKNOWLEDGEMENTS

Financial support was granted from the Norrbacka-Eugenia Foundation, the Frimurare-Barnhuset Foundation, the Foundation for Medical Research at örebro University Hospital (Nyckeln), the Research Unit at örebro University Hospital, and the Research Committee of örebro County Council.

REFERENCES

1. Hermansson LM, Fisher AG, Bernspång B, Eliasson A-C. Assessment of capacity for myoelectric control: a new Rasch-built measure of prosthetic hand control. J Rehabil Med 2005; 37: 166–171.
2. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington: American Educational Research Association; 1999.
3. Krebs DE. Conclusions and future considerations. In: Krebs D, ed. Prehension assessment: prosthetic therapy for the upper-limb child amputee. Thorofare: Slack Incorporated; 1987, pp. 45–48.
4. Pruitt SD, Varni JW, Setoguchi Y. Functional status in children with limb deficiency: development and initial validation of an outcome measure. Arch Phys Med Rehabil 1996; 77: 1233–1238.
5. Wright FW, Hubbard S, Jutai J, Naumann S. The Prosthetic Upper Extremity Functional Index: development and reliability testing of a new functional status questionnaire for children who use upper extremity prostheses. J Hand Ther 2001; 14: 91–104.
6. Hubbard S, Galway HR, Milner M. Myoelectric training methods for the preschool child with congenital below-elbow amputation. J Bone Joint Surg Br 1985; 67-B: 273–277.
7. Wright BD, Stone MH. Best test design. Chicago: MESA; 1979.
8. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2001.
9. Hermansson LM. Assessment of capacity for myoelectric control-en metod för att bedöma myoelektrisk kontroll vid utförande av dagliga sysslor. [A method for assessment of myoelectric control in daily life tasks performance]. örebro: Universitetssjukhuset örebro; 2000 (in Swedish).
10. Wright BD, Masters GN. Rating scale analysis. Chicago: Mesa Press; 1982.
11. Fischer GH. Rasch modeling. In: Everitt BS, Howell DC, eds. Encyclopedia of statistics in behavioral science. Vol 4. Chichester: Wiley; 2005, pp. 1691–1698.
12. Arnould C, Penta M, Renders A, Thonnard J-L. ABILHAND-Kids. A measure of manual ability in children with cerebral palsy. Neurology 2004; 63: 1045–1052.
13. Fisher AG. Assessment of motor and process skills, 3rd edn. Fort Collins, CO: Three Star Press; 1999.
14. Krumlinde-Sundholm L, Eliasson A-C. Development of the Assisting Hand Assessment: a Rasch-built measure intended for children with unilateral upper limb impairments. Sc J Occup Ther 2003; 10: 16–26.
15. Fleiss JL, Levin B, Cho Paik M. Statistical methods for rates and proportions, 3rd edn. Columbia: Wiley-Interscience; 2003, pp. 598–626.
16. Linacre JM. Many-facet Rasch measurement. Chicago: MESA Press; 1994.
17. Linacre JM. A user's guide to Facets Rasch measurement computer program [monograph on the internet]. Chicago: www.Winsteps.com; 1991–2005 [cited 2005 August 18]. Available from: http://www.winsteps.com/facetman/index.htm
18. Lunz ME, Stahl JA. The effect of rater severity on person ability measure: a Rasch model analysis. Am J Occup Ther 1993; 47: 311–317.
19. Bernspång B. Rater calibration stability for the Assessment of Motor and Process Skills. Scand J Occup Ther 1999; 6: 101–109.
20. Bland JM, Altman DG. Statistical methods for assessing agreement between 2 methods of clinical measurement. Lancet 1986; 8: 307–310.