Inter-observer reliability of the “Assessment of Motor Repertoire — 3 to 5 Months” based on video recordings of infants

Toril Fjørtoft a,d,*, Christa Einspieler b, Lars Adde a,c, Liv Inger Strand d

a Department of Clinical Services, Physiotherapy Section, Trondheim University Hospital, Trondheim, Norway
b Institute of Physiology, Center for Physiological Medicine, Medical University of Graz, Austria
c Department of Laboratory Medicine, Children and Women’s Health, Faculty of Medicine, Norwegian University of Science and Technology, Norway
d Department of Public Health and Primary Health Care, Section for Physiotherapy Science, University of Bergen, Norway

ABSTRACT

Objective: A detailed analysis of infant motor behaviour can show up indicators for later neurological impairment. The “Assessment of Motor Repertoire — 3 to 5 Months”, which is part of Prechtl’s general movement assessment, could potentially be used for this purpose. The aim of the present study was to investigate inter-observer reliability in this instrument.

Method: Video recordings of 24 infants (corrected ages 3 to 5 months, gestational ages 24 to 42 weeks) were analysed by four observers. Kappa and ICC statistics were applied in the reliability analysis.

Results: High to very high inter-observer reliability was found in the assessment of “Fidgety Movements” (kappa 0.75–0.91). Agreement on the “Movement Character” was also high (kappa 0.54–0.84), while the assessment of the “Posture” showed the lowest inter-observer reliability (kappa 0.39–0.56). Moderate to high inter-observer reliability (kappa 0.51–0.84) was achieved in the field “Quality of Other Movements”, and moderate in “Repertoire of Co-Existential Other Movements” (kappa 0.51–0.69).

Inter-observer reliability in the assessment of the total “Motor Optimality Score” was very high between all four observers as intraclass correlation coefficient (2,1) was 0.87, and ICCs for the pairwise analyses ranged between 0.80 and 0.94.

Conclusion: Inter-observer reliability in the “Assessment of Motor Repertoire — 3 to 5 Months” was satisfactory in respect of the subcategories and in case of high and low total optimality scores in pairwise assessments. In the total optimality scores, however, there was some inconsistency in the middle range of the scale.

© 2008 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Preterm infants are susceptible to significant risk factors for abnormal neurological outcome [1–3], and perinatal complications [2,4–8] can have a strong influence on a child’s neurological development. There is an urgent need to develop and improve diagnostic tools for an early detection of neurological impairment such as changes in muscle tone or abnormal reflexes [18]. The observation of spontaneous movements in normal foetuses has led to a systematic classification of the children’s developmental stages with those of the average neonates and infants has shown a set of standard movements for each respective age group [20], some of which are described as general movement Assessment (GMA). GMs have been found an effective point of reference for the functional assessment of the developing nervous system [21]. Accordingly, a method for the evaluation of general movements has been developed, known by the term General Movement Assessment (GMA) [21–23]. The GMA has frequently been used in studies for the prognosis of neurological outcomes [24–27]. Studies of preterm and term newborns as well as young infants have shown
that abnormal general movements in preterm infants, abnormal writhing movements and/or the absence of fidgety movements can be related both to brain lesions and to an unfavourable neurological outcome [8,21,27–30].

In the classical GMA the quality of three main periods of general movements is assessed by means of video recordings: preterm general movements, writhing movements and fidgety movements. The analysis is complex and requires a lot of clinical experience [21]. Inter-observer agreement in GMAs has been studied in several groups of infants, agreement being expressed in terms of kappa [21,31,32] or percent [21].

Fidgety movements constitute the characteristic general motor pattern in 3 to 5 month-old infants. They are small movements of moderate speed and variable acceleration of neck, trunk and limbs in any direction, continual in the awake infant, except during fussing and crying [27]. Absence of fidgety movements is considered predictive for later development of cerebral palsy [8,21,27,28,33,34] while normal fidgety movements have been found predictive of normal neurological development [27,29,35]. The GMA has been optimised to improve its predictive value for minor motor impairment and possible cognitive disturbances. Indication of a positive prediction of the GMA for complex minor neurological dysfunction has been reported in several studies [24,32,36–39]. In one study [35], however, the diagnosis of “poor repertoire” – the most frequently observed abnormal GM-pattern in preterm infants – failed to be predictive for the neurological outcome.

After a standardised basic course over five days, 800 observers performed 9000 GMAs in total. Correct agreement with the gold standard was achieved in 83% of the assessments – a result that was improved to 88% after an advanced course [40]. Repeated assessments of 20 GM recordings, carried out after a time interval of 2 years, resulted in a test–retest reliability of 100% for the global judgement [21] and 85% for a more detailed analysis of movement quality based on the same principle of optimality as the “Assessment of Motor Repertoire – 3 to 5 Months” [41].

The global GMA is not suitable for an evaluation of therapeutic effects, which necessitates a detailed assessment. In order to introduce a more detailed approach during the age of fidgety movements, the “Assessment of Motor Repertoire – 3 to 5 Months” [22] has been developed. Based on the optimality concept, this assessment tool [42] places emphasis on finding the best possible condition rather than finding normality, abnormality or pathology. It includes assessment of movements that co-occur with fidgety movements, namely wiggling–oscillating arm movements, swipes, mutual manipulation of fingers, fiddling with clothing, leg lifts, trunk rotation and axial rolling [22].

It takes sufficient inter-observer reliability in order for a different group of testers to use an instrument for scientific and clinical purposes. Before this study, the “Assessment of Motor Repertoire – 3 to 5 Months” had not been subjected to an examination of inter-observer reliability. The aim of this study, which was based on video-recordings of infants, was to determine inter-observer reliability of the above mentioned assessment tool.

2. Subjects and methods

2.1. Design

To determine the degree of inter-observer agreement, a cross-sectional study design was chosen. Four participants (observers A, B, C and D) analysed the same 24 videotapes of infants at the same time, applying the “Assessment of Motor Repertoire – 3 to 5 Months” and following a standardised assessment procedure [21].

2.2. Observers

Before the actual study, the four observers had participated in a four-day basic and a four-day advanced training course on the assessment of GMs [21]; they all had previously used GMA as a diagnostic tool in clinical practice. They were labelled by the characters A to D. Observer A, having accomplished the development of the “Assessment of Motor Repertoire – 3 to 5 Months” as a tool, was highly qualified in the assessment of general movements [22]. Observers B, C and D were highly qualified child physiotherapists. Since only observer A was conversant with the given scoring system, all four observers had completed a joint training workshop before commencing the study. The workshop had consisted of four theoretical lectures and one training session in which ten video recordings of infants had been analysed; one of the four observers had recorded and edited the 16 videos while another had recorded and edited 9 videos in accordance with Precht’s method [21]. In the actual study, the observers were not familiar with the children’s histories – except for observer D, who recognised five children from a previous clinical study [28].

2.3. Subjects

Prior to the study, video recordings of 25 infants aged 3 to 5 months post-term were arranged. The recordings had been carried out at the Department of Paediatrics, Trondheim University Hospital between 1999 and 2005. The Regional Ethics Committee approved the study, and all parents gave their written informed consent, allowing the video recordings to be used for research purposes. The intention was to select a diversified group of children both regarding gestational age and the risk for later neurological impairments. All 25 infants participated in a follow-up programme for children with significant risk factors. Sixteen of the infants had previously participated in a follow-up study [28]. Additional video recordings of 9 infants were selected for the present study from clinical files.

A broad spectrum of infants of various gestational ages and full term infants with various risk factors for later neurological impairments – 13 females and 12 males in total – participated in the study. Birth weight ranged from 680 to 4725 g. Gestational age was 24 to 28 weeks in 13 infants and 29 to 33 weeks in 5 infants; 6 infants had been born at term. Nine infants had shown abnormal ultrasound imaging or MRI findings during their first three months of life (intraventricular haemorrhages or infarcts). Moderate or severe asphyxia had been recorded in 8 infants; 5 infants had been treated for septicaemia during the first four weeks of life. All 25 infants showed peri- and/or neonatal risk factors for later development of neurological problems. One recording had to be discarded, because the child’s motor behaviour did not meet the criteria for assessment [21].

2.4. Video recordings

In compliance with a procedure described by Einspieler et al., representative sequences of movements were selected from the video recordings [23]: The infants were always videoed in supine position for 5 to 10 min and had to be fully awake. Sequences that included crying and fussing were discarded. Accordingly, a total of 24 infants were included in the study; one was discarded, because the video recording did not meet the criteria for assessment.

The average time it took the observers to assess one video (out of 24) was 4.5 min, always ranging between 2 and 5.5 min. Twelve recordings were seen twice, the other 12 three times. In case of two infants, observer A regarded the subcategory “Fidgety Movements” as not assessable and chose not to valuate them. Consequently, only 22 recordings were included in the calculations of kappa values and ICCs for observer A, whereas the other three observers had analysed 24 recordings.

2.5. The assessment tool

“Assessment of Motor Repertoire – 3 to 5 Months” [22] is a tool designed for the assessment of video recordings of infants. It
consists of three main fields of observation: “Movement Patterns” (consisting of 33 items), “Postural Patterns” (13 items), and “Movement Character” (12 items). The overall result (58 items) is taken as a basis for the “Motor Optimality List”, based on the scoring of five subcategories, the first of which rates “Fidgety Movements” as normal (12 points), abnormal (4 points) or absent (1 point); the second subcategory, “Repertoire of Co-Existent Other Movements”, is classified as age-adequate (4 points), reduced (2 points) or absent (1 point); the third subcategory, “Quality of Other Movements” is evaluated by the number of normal or abnormal items within the field “Movement Patterns”: a number of normal patterns (N) higher than that of abnormal patterns (A) scores 4 points; N = A scores 2 points; N < A scores 1 point. The fourth subcategory, “Posture”, is assessed in the same way, based on the items of the second main field of observation, “Postural Pattern”. The fifth subcategory, “Movement Character”, describes the overall movement character observed in all movement categories: smooth and fluent (4 points); abnormal, but not cramped-synchronised (2 points); abnormal and cramped-synchronised (1 point). Finally, adding up the scores of each subcategory results in a total of 5 to 28 points — the “Motor Optimality Score”.

The author has omitted two items from the assessment tool after its first publication [22]: “Saccadic Arm Movements” were not taken into consideration in the present study because their description was insufficient and they could have been confused with abnormal fidgety movements; the category “Mouth Movements” was withdrawn, because, if abnormal, they co-occur with abnormal “Tongue Movements”. “Hand–Face Contact” and “Hand–Mouth Contact” were regarded as one item. These changes, however, did not affect the total optimality score. The subcategories “Fidgety Movements”, “Repertoire of Co-Existent Other Movements”, “Posture” and “Movement Character” were all given numeric values as a result of a sum of nominal values. These numeric values added up to a total “Motor Optimality Score”.

Fig. 1. Inter-tester reliability of the instrument “Assessment of Motor Repertoire — 3 to 5 Months”: pair wise correlations of test results and linear regression lines for the total “Motor Optimality Score” by observer A, B, C and D.
2.6. Assessment procedure

The assessment of the 24 video recordings was performed in the same room by all observers, using a large video screen. There was no possibility for the observers to communicate. Upon request, they were allowed to view the video sequences repeatedly. Each observer saw each video recording the same number of times and for the same length of time. The time that was spent on each infant was recorded. The scoring sheets were numbered consecutively from 1 to 24 in analogy to the infants.

2.7. Statistics

SPSS version 14.0 was used for statistical analyses. In the five subcategories of the assessment tool, the degree of inter-observer agreement was identified by means of kappa statistics or expressed in terms of percent agreement if the kappa value could not be determined, and it was arranged on an ordinal scale. Cohen's kappa is a statistical measure that is used to determine inter-observer agreement, taking into account agreement by chance [43]. The results were interpreted according to guidelines adapted from Landis and Koch [44], who classify a κ value of <0.20 as poor agreement, of 0.21–0.40 as fair, of 0.41–0.60 as moderate, of 0.61–0.80 as good, and of 0.81–1.00 as very good agreement.

Intraclass correlation coefficient (ICC) statistics was applied to examine pairwise agreement of sum scores among the observers. ICCs are correlation coefficients that allow comparison of two or more repeated measurements; the method is based on the repeated measures analysis of variance (ANOVA) [43]. For the "Motor Optimality Score", ICC (2,1) statistics was applied to examine pairwise inter-observer agreement (A–B, A–C, A–D, B–C, B–D, C–D), and agreement among all four observers (A–B–C–D). ICC (2,1) was chosen so the result could be generalised to other observers [45]. The measurement error was termed "Sw"; it was calculated as the square root of the mean within-subject variance. The difference between an observer's evaluation of an infant and the true value was expected to be less than 1.96 Sw in 95% of the observations [46].

3. Results

By tendency, the children that participated in the study either ranged at the lower end or at the head of the 5- to 28-point total "Motor Optimality Score" (Fig. 1). Inter-observer agreement for the total score – expressed in terms of ICC (2,1) values – was high, as is shown in Table 1. Regarding pairwise agreement, ICC (2,1) values ranged between 0.80 and 0.95. Pairwise correlations between the observers are shown in scatter plots (Fig. 1). Overall inter-observer agreement was 0.87.

The measurement error (Sw) between the various pairs of observers in the assessment of the "Motor Optimality Score" ranged from 2.42 to 4.25. Variability among the observers was found to be high in case of children who scored in the middle range of the scale. The overall Sw between the observers was 3.47, which implies that in

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-tester reliability of the total &quot;Motor Optimality Score&quot; in pair wise between four observers (A–D) and for all observers, reporting Intra Class Correlation Coefficient (ICC) and within subject standard deviation (Sw).</td>
</tr>
<tr>
<td>Observers</td>
</tr>
<tr>
<td>A–B</td>
</tr>
<tr>
<td>A–C</td>
</tr>
<tr>
<td>A–D</td>
</tr>
<tr>
<td>B–C</td>
</tr>
<tr>
<td>B–D</td>
</tr>
<tr>
<td>C–D</td>
</tr>
<tr>
<td>A–B–C–D</td>
</tr>
</tbody>
</table>

N = 24 observations for B, C, and D, and 22 for A.

95% of the cases the measurement error will be within ±3.47×1.96, which explains 13.6 points on the 5- to 28-point optimality score scale. In the subcategory "Fidgety Movements", kappa values could only be calculated for three pairs of observers: A–B=0.91, A–C=0.82 and B–C=0.75 (Table 2). Since only observer D attested that children numbers 3 and 24 showed abnormal fidgety movements, no symmetric 2-way table could be constructed, and consequently no kappa value for fidgety movements could be calculated between observer D and the other observers. Therefore, agreement between observers A–D, B–D and C–D regarding the subcategory "Fidgety Movements" was expressed in terms of percent – 82%, 75% and 88% respectively. Observer A considered children 19 and 24 not to meet the criteria for an assessment of fidgety movements and therefore scored no fidgety movements for them.

In the other subcategories, data from all 24 infants were included in the analysis. Moderate inter-observer reliability was achieved in the assessment of "Repertoire of Co-Existent Other Movements", with kappa values ranging between 0.48 and 0.69 and one single value under 0.5. Regarding the "Quality of Other Movements", inter-observer reliability was moderate to high, with kappa values ranging from 0.51 to 0.84 and three out of six values higher than 0.6. The assessment of "Posture" resulted in moderate kappa values, ranging from 0.39 to 0.56 with only two values above 0.5. Movement Character appears to be the subcategory easiest to assess, since here the results were most consistent: Kappa values ranged between 0.54 and 0.84, with five values above 0.60 (Table 2).

Observer D may have recognised five video recordings from a previous study [28], but it must be added that none of observer D's present scores differed significantly from the other observers' scores for these videos.

4. Discussion

The objective of this study was to investigate inter-observer reliability in the "Assessment of Motor Repertoire – 3 to 5 Months". Four observers qualified and then participated in the assessment of video recordings of spontaneous movements in a large number of infants. ICCs between 0.80 and 0.93 for the "Total Optimality Score" indicate high to very high reliability according to Munro's descriptive terms of the intraclass correlation coefficient [46]. But regardless of the fact that ICCs were high, the great variance of the scores made it difficult to interpret the results. A look at the scatter plots (Fig. 1) reveals that most scores were located either at the upper or at the lower end of the 5- to 28-point scale. Such a broad range of scores may result in artificially high ICC values; but then the observers seemed to agree both on the respective high scores and on the low scores in the category "Total Optimality Score". Those few valuations in the middle range of the scale showed large variability, and the overall within-subject standard deviation was wide. Consequently, it was difficult to determine inter-observer reliability for the middle range of the "Total Optimality Score" on the basis of the present study.
It must be taken into consideration that, within the assessment of the “Motor Optimality Score”, the subcategory “Fidgety Movements” accounted for as much as 12 out of 28 points. Thus, the assessment of “Fidgety Movements” – which itself showed good inter-observer agreement [21] – had a significant effect on the ICCs for the total “Motor Optimality Score”.

The points achieved in the subcategories “Quality of Other Movements” and “Posture” were calculated on the basis of 33 or 13 items respectively. The observed patterns were described either as normal or as abnormal, with the total points achieved per subcategory being the sum of all respective normal and abnormal observations. Accordingly, the result was not simply based on the inter-observer agreement in each item. It is only the sum that counts. Therefore, there might be a certain degree of expected chance agreement involved in these subcategories, which was not examined further in this study. Even if inter-observer agreement on each item of these subcategories turned out to be low, the points achieved for “Quality of Other Movements” and “Posture” came out with high agreement, which again influenced the “Motor Optimality Score” and ultimately the ICC values of reliability.

In the subcategory “Fidgety Movements”, observers A–B, A–C and B–C achieved high or very high agreement, expressed in terms of kappa values. These results corresponded with previous findings, which show that inter-observer agreement in the assessment of fidgety movements is rather high in general [40]. From a clinical point of view, these findings are of utmost importance as the presence or absence of fidgety movements has a high prognostic value [27]. Those three pairwise observations in which agreement was expressed in percent without taking into consideration agreement by chance are harder to interpret. A percentage of 75% to 88% would seem satisfactory as it clearly exceeds potential agreement by chance. Six pairwise agreements were carried out for the other subcategories, agreement being expressed by means of kappa values. In the subcategory “Repertoire of Co-Existant Other Movements”, moderate inter-observer reliability was achieved in five pairwise observations and high interobserver reliability in one pairwise observation. Regarding “Quality of Other Movements” and “Movement Character”, inter-observer agreement was also moderate to high in all pairwise observations. Yet in the assessment of the subcategory “Posture” it proved more difficult to achieve high inter-observer agreement than in the other categories, since here, four kappa values ranged between 0.39 and 0.48, the other two being 0.54 and 0.56. It has been argued, however, that in studies that apply observational methods, lower reliability values should be acceptable than in studies that use more objective methods of measurement [43]. Taking into consideration that the present study was based on visual observations and clinical judgement, the kappa value for “Posture” might be regarded satisfactory [43].

In the present study we preferred Cohen's kappa statistics to percent agreement in order to examine inter-observer agreement on ordinal data. This sort of analysis has also been used in a number of previous studies on this instrument aimed at, the results should be interpreted with due care, since the observations do not cover the optimality scale sufficiently and the measurement error – 59% of the total score of all observers taken together (A–B–C–D) – was found to be rather high. Those few children who ranged in the middle section of the scale contributed substantially to the variability in scores among the testers. Another survey of the characteristics of the children tested was conducted to examine possible reasons for the fact that the testers had scored so inconsistently. Five recordings with a total optimality score divergence of more than 12 points between two testers were identified and reanalysed. In all five recordings, the discrepancy was located in diverging scores for fidgety movements. In general, these children moved less and seemed to be partly distracted by staff and equipment – which in turn may have influenced the observers' judgements. Perhaps this indicates that the recording conditions were not always ideal for data acquisition. In order to obtain good video quality and high inter-observer agreement it is of paramount importance that the described procedure be followed carefully when selecting the recordings.

5. Conclusion

The present study on inter-observer agreement in the “Assessment of Motor Repertoire — 3 to 5 Months” has produced satisfactory kappa values for the subcategories and high ICC values for the total score. The subcategory “Fidgety Movements” showed high to very high inter-observer agreement across the 6 pair-wise analyses, while there was less agreement in the other subcategories, ranging between moderate and high. The reliability based on ICC values was hard to interpret since the scores were clustered mainly around the upper and lower ends of the optimality scale. Regarding the total scores, there was great variability in the middle range of the scale. Reanalyses of five of the recordings indicated that this variability was due to inconsistent judgement of fidgety movements. Further studies are needed to examine reliability of the scale — including scores along the whole scale.

Acknowledgements

This study was supported by the Department of Clinical Services, Trondheim University Hospital, in Trondheim, and by grants from the Norwegian Fund for Post Graduate Training in Physiotherapy. The first author would like to thank the physiotherapist PhD Gunn Kristin Øberg, Tromsø University, Norway, for her kind help in the assessment of the video recordings.

References


