Early Human Development 87 (2011) 633–639

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

Inter- and intra-observer agreement of Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants

I. Bernhardt [a,*], M. Marbacher [a], R. Hilfiker [b], L. Radlinger [b]

[a] Physiotherapy Institute, Inselspital, Bern University Hospital, Switzerland
[b] Bern University of Applied Sciences, Health, Research & Development Physiotherapy, Bern, Switzerland

ARTICLE INFO

Article history:
Received 25 June 2010
Received in revised form 4 March 2011
Accepted 28 April 2011

Keywords:
Physiotherapy
Pediatrics
Prechtl
General movements
Reliability

ABSTRACT

Background: Prechtl's method on the qualitative assessment of general movements (GMs) has been shown to be a good predictor of neurological outcome. There is substantial evidence that this method has good inter- and intra-observer agreement.
Aims: We wanted to find out whether this high agreement is reproducible in the clinical setting.
Study design: Reliability study (inter- and intra-observer agreement).
Subjects: Twenty video-sequences of children at the age of preterm and writhing movements (31–41 weeks postmenstrual age) and 10 video-sequences of children at the fidgety movements age (49–54 weeks postmenstrual age) were rated by five physiotherapists.
Outcome measures: Intra- and inter-observer agreements were analyzed with percentage agreement and with nominal kappa statistic with bootstrapped bias corrected 95% confidence intervals.
Results: We found fair to substantial inter-observer reliability for the six response categories (time-point 1 (t1): median kappa 0.44, range 0.27 to 0.59, time-point 2 (t2): median kappa 0.55, range 0.41 to 0.77) and fair to almost perfect for the normal/abnormal ratings (t1: median kappa 0.53, range 0.29 to 0.61, t2: median kappa 0.63, range 0.29 to 0.85). There was statistically significant improvement from t1 to t2 for the six response categories. The intra-observer reliability for the 9-week interval was moderate to almost perfect (median kappa 0.68, range 0.41 to 0.86).
Conclusions: We were not able to exactly reproduce the generally very good results. In our clinical setting now videos are evaluated by at least two trained therapists and the results are discussed, if necessary, to reach a consensus.

1. Introduction

Preterm and high-risk infants (e.g., low birth weight) have a higher risk of an abnormal neuro-developmental outcome [1,2]. Early detection of motor impairments is important for a timely start of intervention programs [3–5]. Prechtl's method on the qualitative assessment of general movements (GMs) has proven to be an objective, reliable and valid diagnostic tool for early recognition of brain dysfunction [6–15]. It was developed by Prechtl and co-workers. They founded in 1997 the General Movements Trust (GM Trust) in order to promote Prechtl's method. Since then the GM assessment is increasingly and widely employed in clinical routine and supported by the licensed instructors of GM Trust, who educate physiotherapists and physicians in their training courses and educations concerning the detection of movement patterns [16]. Human fetuses, preterm, term and young infants during the first months of life, have a repertoire of distinct centrally generated movement patterns. Among these different spontaneous movement patterns, the GMs are the most effective to evaluate the integrity of the young nervous system [17]. They can be observed from 9 weeks postmenstrual age (PMA) up until 5 months post-term age (PTA) [18]. The quality of these movements changes if the nervous system is impaired [19]. GMs are age-specific. Before term age they are called fetal and preterm GMs, around term age writhing movements and from 9 weeks post-term age onwards fidgety movements. Preterm and writhing GMs involve the whole body and they are characterized by the variable sequence of arm, leg, neck and trunk movements. Their intensity, force and speed begin and end gradually. Frequent changes in direction and rotations along the axis of the limbs make the movements fluent, elegant, complex and variable [16]. Abnormal patterns of GMs during the preterm and writhing periods are poor repertoire (PR), cramped synchronized (CS) and chaotic (Ch). The character of GMs changes to fidgety movements (FMs) around 9 weeks post-term. FMs are small, circular movements of moderate speed and variable acceleration of neck, trunk and limbs in all directions. Abnormal patterns of GMs during the fidgety movement period are rated as absent (F−) or

* Corresponding author at: Physiotherapy Institute, Inselspital, Bern University Hospital, 3010 Bern, Switzerland. Tel.: +41 31 632 94 58; fax: +41 31 632 15 72.
E-mail address: iris.bernhardt@insel.ch (I. Bernhardt).

0378-3782/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.earlhumdev.2011.04.017

abnormal (AF) fidgety movements. If the normal quality of GMs has changed this is a reliable indicator of brain dysfunction. The quality of GMs is assessed by means of video recordings and is based on the visual Gestalt perception of the observer [13]. Gestalt effect refers to the visual recognition of figures and whole forms instead of a collection of details [20]. In the physiotherapy unit of the Children's Hospital, Bern University Hospital, GM is used for standard assessment of preterms born before the 28th week of gestation, after severe asphyxia or with neurological abnormality in development. In the clinical field it is very important to be able to apply a method that is valid, reliable, non-invasive, low-cost, non-time-consuming and highly predictive for a better identification of infants that can profit from early intervention services [3,4,21,22]. It is claimed that the assessment can be reliably performed by trained physiotherapists. There is substantial evidence that Prechtl's method has good inter- and intra-observer agreement [12–15,23,24]. But most studies have been accomplished by members of the GM Trust who are highly skilled in evaluating GMs. We wanted to find out what is the reliability of the inter- and intra-observer agreement in a clinical setting. The specific research questions were (a) whether the inter-observer agreement is acceptably high (at least classified as substantial, median kappa>0.6) between five experienced physiotherapists, (b) whether the intra-observer reliability is acceptably high (median kappa>0.6) between two rating sessions with an interval of 9 weeks, and (c) whether there is a difference in the agreement if the videos are classified using the six different GM categories (nominal response categories) or only normal/abnormal (dichotomous).

2. Subjects and Methods

2.1. Study Design

In this psychometric study, five physiotherapists rated the quality of the movements of 30 children on existing video-sequences at two time-points.

2.2. Subjects

To test inter- and intra-observer agreement, 20 video-sequences from 20 children at the preterm and writhing movements age and 10 video-sequences of children at the fidgety movements age were chosen (Table 1). All parents signed an informed consent form.

The videos of fidgety movements were randomly selected from the archive of routinely filmed children, each video lasted 2 minutes. The videos for the preterm and writhing movements were randomly selected from an external archive, the duration of the videos ranged from 31 seconds to 2.56 minutes. The duration of the videos was quite short but short sequences in test–retest do not influence reliability. But it is obvious that in a clinical setting one might choose at least three such representative GM sequences to use Prechtl's method as a valid assessment. The selection of all videos was done by an independent person who was not involved in the study and who had no specific training or knowledge about general movements. The videos were selected on the basis of inclusion and exclusion criteria. Inclusion criteria were correct age group, adequate behavioral state and correct recording, which includes behavioral state 4 (eyes open, not crying, irregular respiration, movements present) or state 2 (REM-sleep, superficial sleep, eyes closed, unsteady breathing, moving) for preterm and writhing movements [25]. Fidgety movements should be rated only with children awake in state 4. Recordings have to be made from above, sagittal view with the face clearly visible. The child had to lie on his back although turning on his side was allowed and wearing only diapers. The underlay had to be unicolored.

Exclusion criteria were bedding material like pillows or sheets, crying, whining, hiccupping or having a pacifier, toys above or in the bed and external stimulation by caregivers [16].

Table 1
Number of videos in each GM group and age of children.

Age group	Videos (no.)	Post-menstrual age	GM quality
1	3	31–33 weeks	Preterm movements
2	9	34–36 weeks	Preterm movements
3	8	37–41 weeks	Writhing movements
4	10	49–54 weeks	Fidgety movements

2.3. Observers

The five observers (A, B, C, D, E) had successfully participated in a 4-day basic course (C, D, E in 2003; B in 2005; A in 2007) and two of them had also taken part in a 4-day advanced training course (E in 2006, B in 2007) of GMs. So they had several years of experience in rating children by this method in clinical practice. Before starting the study, the five observers discussed the typical criteria for GMs for each classification, created an information leaflet based on these discussions and watched the educational DVD with examples of each age group [16]. In the actual study the observers were blind to the clinical history details (i.e., gender, birth weight, cranial ultrasound).

2.4. Procedure

All five observers watched the 30 videos together in one room. The videos were projected on the wall and each video was shown twice in a row. At rating time-point 1 (A1, B1, C1, D1, E1), 20 recordings (preterm and writhing movements) were shown with a 1-hour break after 10 videos. Ahead of each age group, the observers watched an educational video example with normal quality of general movements for this specific age group. The 10 videos with the fidgety movements were shown 1 week later. Each person rated the movements individually. Talking or showing ratings was not allowed. Ratings had to be distinct. All observers used the same standardized rating sheet, which was used also during the training courses, with the indications of the age of the child, the case number and the different rating possibilities (N, PR, Ch, CS, AF, F−). All observers had the information leaflet, which was created before starting the study, with the criteria for each classification. At rating time-point 2 (A2, B2, C2, D2, E2), 9 weeks later, the videos were shown again in the same order and setting.

After 6 months, the observers scored all videos again but at this time they exchanged their evaluations and discussed the results in order to reach a consensus.

2.5. Statistical Analysis

For the sample size calculation, we expected the kappa values to be above 0.9 and the proportion of abnormal to normal ratings as 50:50. To achieve a power of 80% with a significance level of 5% to test against a lower limitation of the kappa values of 0.41 (as the lowest value considered moderate agreement [26]), we chose a sample size of 30 videos [27].

Intra-observer and inter-observer agreements were analyzed with percentage agreement and with nominal kappa statistics with bootstrapped bias corrected 95% confidence intervals. We analyzed each combination of the five observers at rating time-point 1 and rating time-point 2 (intra-observer) and the agreement for each observer between rating time-point 1 and rating time-point 2. To analyze whether there was a significant change in the agreements between the rating time-points, we used a Wilcoxon test (exact) for the paired data (significance level (α) set at 0.05).

Furthermore, we calculated one overall kappa value for rating time-points 1 and one for rating time-points 2.

We present two analyses: first we analyzed the videos of the preterm, writhing and fidgety movements together. In a second analysis, we calculated the kappa values for preterm and writhing movements together, and the kappa values for the ratings of the fidgety movements separately. We distinguished the six nominal response category ratings (N, PR, Ch, CS, AF, F−, no movement was rated chaotic) and the dichotomous ratings (normal/abnormal).

Kappa values were calculated with the command kap (STATA, Version 9.2, StataCorp, College Station, Texas, USA) and the user-written STATA-command kapci for the bias-corrected bootstrap confidence intervals [28]. These STATA-commands allow calculations for non-symmetric tables (i.e., when one or more rating categories are used by one observer but not by the other).

According to Landis and Koch [26], kappa values can be classified as follows: below zero = poor, zero to 0.20 = slight, 0.21 to 0.4 = fair, 0.41 to 0.6 = moderate, 0.61 to 0.8 = substantial, 0.81 to 1 = almost perfect.

3. Results

3.1. Overall Ratings

Table 2 shows the ratings of the 30 videos at the two different rating time-points by the five observers. The last column shows the consensus ratings after 6 months. There was no consensus for three videos. No movement on the videos was rated chaotic.

The kappa value for the overall agreement at rating time-point 1 was 0.42 ($p<0.0001$) for the six nominal response category (N, PR, Ch, CS, AF, F−) ratings and 0.48 ($p<0.0001$) for the normal/abnormal ratings.

At rating time-point 2, the overall kappa value was 0.56 for the six response categories and 0.60 for the normal/abnormal ratings.

3.2. Rating for Individual Pairs of Observers

At rating time-point 1, the median kappa values for the inter-observer agreement were 0.44 (range between 0.27 and 0.59) for the six response category ratings and 0.53 (range between 0.29 and 0.61) for the normal/abnormal ratings (Table 3).

At rating time-point 2, 9 weeks later, the median kappa value for the inter-observer agreement was 0.55 (range 0.41 to 0.77) for the six response category ratings and 0.63 (range 0.29 to 0.85) for the normal/abnormal ratings.

There was an improvement in the ratings between the first and the second rating time-point (Fig. 1), the improvement was significant for the six response categories (for the percentages: $p=0.006$, for kappa values: $p=0.004$). The improvement was not statistically significant for the normal/abnormal ratings (for percentages: $p=0.053$, for kappa values: $p=0.105$) (2-tailed, exact p-values from the Wilcoxon test).

3.3. Intra-observer Agreement

The median kappa values for the intra-observer agreement were 0.68 (range between 0.41 and 0.86) for the six response categories ratings and 0.77 (range between 0.40 and 0.86) for the normal/abnormal ratings (Table 3).

3.4. Difference in Agreement Between Nominal and Dichotomous Response Categories

The difference between the kappa values for the six response category ratings and the normal/abnormal ratings did not differ significantly (exact Wilcoxon test, 2-tailed p-value for intra-observer kappa: 0.50, p-value for inter-observer at rating time-point 1: 0.063, p-value for inter-observer at rating time-point 2: 0.192).

Table 2
Ratings of the five observers for the 30 videos.

Case	Age group	Video duration	GM quality (ratings at time-point 1)					GM quality (rating at time-point 2)					GM quality (consensus)
			A1	B1	C1	D1	E1	A2	B2	C2	D2	E2	
1	1	1:42	N	N	N	N	N	N	N	N	N	N	N
2	1	2:28	N	N	PR	N	N	PR	PR	PR	PR	N	N
3	1	1:14	N	N	PR	N	PR	PR	PR	PR	N	PR	N
4	2	1:58	PR	PR	CS	CS	PR	PR	PR	CS	CS	PR	PR
5	2	1:53	N	N	CS	CS	CS	CS	CS	CS	CS	CS	CS
6	2	0:31	N	N	N	N	N	PR	N	PR	N	N	N
7	2	1:45	PR	PR	PR	CS	PR	CS	CS	PR	CS	PR	CS
8	2	2:12	N	N	N	N	N	CS	N	N	N	N	N
9	2	2:08	CS	CS	CS	CS	PR	CS	PR	PR	CS	PR	CS
10	2	2:56	PR	PR	PR	N	PR	PR	PR	PR	N	PR	PR
11	2	1:47	PR	PR	PR	PR	PR	PR	PR	PR	PR	PR	PR
12	2	2:03	CS	PR	PR	CS	PR	CS	PR	PR	PR	PR	PR
13	3	2:25	N	N	N	PR	PR	N	N	N	N	N	N
14	3	1:49	PR	N	PR	CS	CS	CS	CS	CS	CS	CS	CS
15	3	2:05	CS	CS	PR	CS	CS	CS	CS	CS	CS	CS	CS
16	3	2:23	N	PR	N	PR	PR	N	N	N	PR	PR	*N or PR?*
17	3	2:04	CS	N	PR	PR	PR	CS	PR	PR	PR	N	*PR or CS?*
18	3	1:43	PR	N	PR	N	PR	PR	CS	CS	N	PR	CS
19	3	1:44	PR	PR	N	CS	PR	CS	CS	CS	CS	PR	CS
20	3	2:31	PR	PR	PR	CS	PR	PR	PR	PR	PR	PR	PR
21	4	2:00	F−	F−	F−	F−	F−	F−	F−	F−	F−	F−	F−
22	4	2:00	N	N	N	F−	N	N	N	N	F−	N	N
23	4	2:00	N	N	N	N	N	N	N	N	N	N	N
24	4	2:00	N	N	AF	N	F−	F−	N	AF	N	F−	F−
25	4	2:00	N	AF	AF	F−	N	N	AF	AF	F−	N	N
26	4	2:00	N	N	N	N	N	N	N	N	N	N	N
27	4	2:00	N	N	N	N	N	N	N	N	N	N	N
28	4	2:00	F−	F−	AF	F−	AF	N	F−	AF	F−	F−	*No consensus*
29	4	2:00	N	AF	AF	AF	AF	AF	AF	AF	AF	AF	AF
30	4	2:00	F−	N	AF	N	N	N	N	N	N	N	N

N (normal age-specific GMs), PR (poor repertoire), Ch (chaotic), CS (cramped synchronized), AF (abnormal fidgety movements), F− (absence of fidgety movements); A through E: five observers.

Table 3
Percentage agreement and Kappa values for the different observer pairs (inter-observer and intra-observer) for all 30 videos.

Comparison	Nominal response categories			Normal/Abnormal rating		
	%Agreement	Kappa	95% CI	%Agreement	Nominal Kappa	95% CI
Inter-observer rating time-point 1						
A1 vs. B1	73	0.59	0.33 to 0.80	77	0.53	0.22 to 0.80
A1 vs. C1	57	0.40	0.20 to 0.60	77	0.53	0.24 to 0.80
A1 vs. D1	50	0.30	0.08 to 0.55	70	0.40	0.09 to 0.73
A1 vs. E1	60	0.43	0.24 to 0.68	77	0.53	0.22 to 0.80
B1 vs. C1	57	0.39	0.12 to 0.63	67	0.35	0.07 to 0.67
B1 vs. D1	60	0.44	0.25 to 0.67	80	0.61	0.32 to 0.86
B1 vs. E1	67	0.52	0.29 to 0.74	73	0.48	0.20 to 0.80
C1 vs. D1	43	0.27	0.08 to 0.50	67	0.29	0.06 to 0.63
C1 vs. E1	63	0.48	0.23 to 0.71	80	0.55	0.21 to 0.84
D1 vs. E1	57	0.44	0.25 to 0.65	80	0.57	0.26 to 0.86
Inter-observer rating time-point 2						
A2 vs. B2	70	0.59	0.38 to 0.82	83	0.63	0.27 to 0.86
A2 vs. C2	67	0.55	0.34 to 0.78	90	0.77	0.44 to 1.00
A2 vs. D2	57	0.41	0.18 to 0.65	67	0.29	−0.07 to 0.62
A2 vs. E2	67	0.54	0.31 to 0.78	80	0.57	0.25 to 0.86
B2 vs. C2	83	0.77	0.58 to 0.95	93	0.85	0.61 to 1.00
B2 vs. D2	73	0.64	0.41 to 0.82	83	0.65	0.32 to 0.87
B2 vs. E2	73	0.63	0.41 to 0.84	83	0.65	0.32 to 0.87
C2 vs. D2	63	0.52	0.30 to 0.77	77	0.49	0.16 to 0.82
C2 vs. E2	67	0.54	0.33 to 0.76	83	0.64	0.33 to 0.87
D2 vs. E2	60	0.45	0.20 to 0.68	73	0.44	0.10 to 0.75
Intra-observer						
A1 vs. A2	60	0.44	0.22 to 0.69	70	0.40	0.09 to 0.68
B1 vs. B2	67	0.53	0.28 to 0.76	77	0.54	0.24 to 0.84
C1 vs. C2	77	0.68	0.48 to 0.87	90	0.77	0.44 to 1.00
D1 vs. D2	87	0.82	0.62 to 0.95	93	0.86	0.63 to 1.00
E1 vs. E2	90	0.86	0.69 to 1.00	93	0.86	0.63 to 1.00

A1 = observer A at rating time-point 1, B2 = observer B at rating time-point 2, etc.
The nominal response categories were N (normal age-specific GMs), PR (poor repertoire), Ch (chaotic), CS (cramped synchronized), AF (abnormal fidgety movements), F− (absence of fidgety movements).
The dichotomous rating was normal/abnormal.

3.5. Ratings of Preterm and Writhing Movements

The median kappa value for the intra-observer agreement for the four response category ratings (N, PR, Ch, CS) was 0.53 (range 0.25 to 0.82) and for the normal/abnormal ratings 0.74 (range 0.30 to 0.78) (Table 4).

The median kappa for inter-observer agreement at rating time-point 1 was 0.33 (range 0.16 to 0.60) for the four response category ratings and 0.49 (range 0.20 to 0.63) for the normal/abnormal ratings. At rating time-point 2, the median kappa value was 0.47 (range 0.33 to 0.76) for the four response category ratings and 0.50 (range 0.24 to 0.86) for the normal/abnormal ratings.

3.6. Ratings of Fidgety Movements

Table 5 shows the percentage agreements and the kappa values for the ratings of fidgety movements. The median kappa value for intra-observer agreement for the three response category ratings (N, AF, F−) was 0.83 (range 0.11 to 1.00) and for the normal/abnormal ratings 1.00 (range 0.05 to 1.00).

The median kappa for the inter-observer agreements at rating time-point 1 was 0.38 (range 0.23 to 0.67) for the three response category ratings and 0.42 (range 0.20 to 0.80) for the normal/abnormal ratings. At rating time-point 2, the median kappa value was 0.51 (range 0.29 to 0.80) for the three response category ratings and 0.60 (range 0.20 to 0.80) for the normal/abnormal ratings.

Overall, intra-observer agreement was higher for the rating of fidgety movements compared to the ratings of the preterm and writhing movements, but this was not the case for all individual pairs of observers (Fig. 2a and b).

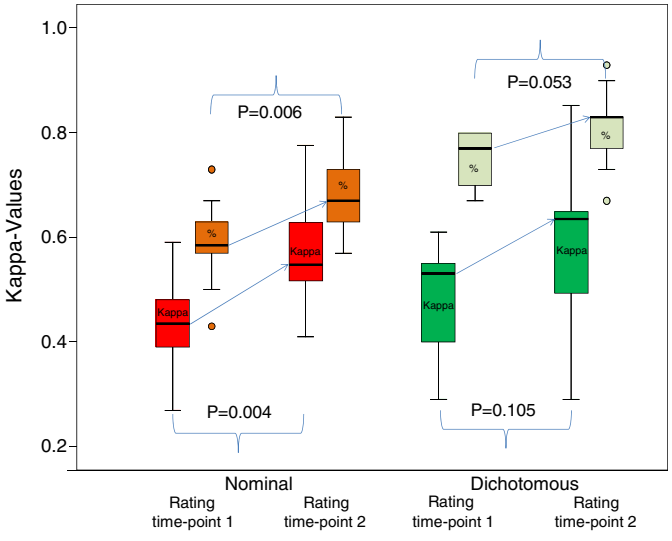

Fig. 1. Boxplots of the percentage agreements and the kappa values of the 10 combinations of the five observers at rating time-point 1 and rating time-point 2 for the nominal N (normal age-specific GMs), PR (poor repertoire), Ch (chaotic), CS (cramped synchronized), AF (abnormal fidgety movements), F− (absence of fidgety movements) response categories. The bold line in the box indicates the median (50th percentile), the lower boundary of the box indicates the 25th percentile and the upper limit of the box the 75th percentile. The box indicates the interquartile range (IQR). The whiskers indicate 1.5 times the IQR, i.e., the lowest or highest value still within 1.5 IQR. The points are extreme values, i.e., between 1.5 and 3 IQR from the boxes away. % = Percentage agreement; kappa = kappa values. The *p*-values are probabilities from a two-tailed, Wilcoxon test (exact).

Table 4
Percentage agreements and Kappa values for preterm and writhing movements (videos 1 to 20).

Comparison	Nominal response categories			Normal/Abnormal rating		
	%Agreement	Kappa	95% CI	%Agreement	Nominal Kappa	95% CI
Inter-observer rating time-point 1						
A1 vs. B1	75	0.60	0.27 to 0.85	80	0.60	0.26 to 0.90
A1 vs. C1	60	0.37	0.06 to 0.68	80	0.57	0.14 to 0.89
A1 vs. D1	45	0.20	−0.09 to 0.48	75	0.47	0.04 to 0.80
A1 vs. E1	60	0.37	0.08 to 0.7	80	0.55	0.22 to 0.90
B1 vs. C1	50	0.19	−0.16 to 0.56	60	0.20	−0.17 to 0.63
B1 vs. D1	50	0.29	0.05 to 0.60	75	0.50	0.17 to 0.89
B1 vs. E1	65	0.44	0.20 to 0.77	70	0.40	0.12 to 0.80
C1 vs. D1	40	0.16	−0.11 to 0.42	65	0.21	−0.24 to 0.68
C1 vs. E1	60	0.29	−0.06 to 0.72	80	0.47	0.00 to 0.89
D1 vs. E1	55	0.39	0.18 to 0.68	85	0.63	0.27 to 1.00
Inter-observer rating time-point 2						
A2 vs. B2	70	0.54	0.21 to 0.84	90	0.69	0.27 to 1.00
A2 vs. C2	65	0.45	0.09 to 0.77	95	0.69	0.27 to 1.00
A2 vs. D2	55	0.33	0.06 to 0.66	70	0.24	−1.10 to 0.69
A2 vs. E2	55	0.33	0.07 to 0.64	75	0.31	−0.10 to 0.74
B2 vs. C2	85	0.76	0.76 to 1.00	95	0.86	0.50 to 1.00
B2 vs. D2	70	0.55	0.27 to 0.85	80	0.53	0.08 to 0.89
B2 vs. E2	70	0.53	0.16 to 0.83	85	0.63	0.18 to 1.00
C2 vs. D2	65	0.48	0.20 to 0.77	75	0.39	0.00 to 0.86
C2 vs. E2	65	0.44	0.10 to 0.77	80	0.47	−0.05 to 0.88
D2 vs. E2	55	0.34	0.00 to 0.67	75	0.43	−0.05 to 0.88
Intra-observer						
A1 vs. A2	60	0.42	0.00 to 0.89	75	0.42	0.00 to 0.89
B1 vs. B2	50	0.25	−0.02 to 0.56	65	0.30	−0.03 to 0.69
C1 vs. C2	70	0.52	0.17 to 0.83	90	0.74	0.34 to 1.00
D1 vs. D2	80	0.70	0.43 to 0.92	90	0.78	0.39 to 1.00
E1 vs. E2	90	0.82	0.55 to 1.00	90	0.74	0.34 to 1.00

A1 = observer A at rating time-point 1, B2 = observer B at rating time-point 2, etc.
The nominal response categories were N (normal age-specific GMs), PR (poor repertoire), Ch (chaotic), CS (cramped synchronized).
The dichotomous rating was normal/abnormal.

Table 5
Percentage agreements and Kappa values fidgety movements (videos 21 to 30).

Comparison	Nominal response categories			Normal/Abnormal rating		
	%Agreement	Kappa	95% CI	%Agreement	Nominal Kappa	95% CI
Inter-observer rating time-point 1						
A1 vs. B1	70	0.42	−0.08 to 1.00	70	0.35	−0.25 to 1.00
A1 vs. C1	50	0.28	0.06 to 0.75	70	0.44	0.14 to 1.00
A1 vs. D1	60	0.25	−0.25 to 0.8	60	0.20	−0.22 to 0.80
A1 vs. E1	60	0.23	−0.14 to 0.78	70	0.35	−0.20 to 1.00
B1 vs. C1	70	0.53	0.19 to 1.00	80	0.62	0.29 to 1.00
B1 vs. D1	80	0.67	0.33 to 1.00	90	0.80	0.55 to 1.00
B1 vs. E1	70	0.46	−0.05 to 1.00	80	0.58	−0.11 to 1.00
C1 vs. D1	50	0.30	0.02 to 0.75	70	0.40	−0.09 to 1.00
C1 vs. E1	70	0.53	0.17 to 1.00	80	0.62	0.29 to 1.00
D1 vs. E1	60	0.33	−0.09 to 0.84	70	0.40	−0.09 to 1.00
Inter-observer rating time-point 2						
A2 vs. B2	70	0.42	−0.13 to 1.00	70	0.35	−0.25 to 1.00
A2 vs. C2	70	0.49	0.14 to 1.00	80	0.60	0.19 to 1.00
A2 vs. D2	60	0.29	−0.32 to 0.83	60	0.20	−0.25 to 0.80
A2 vs. E2	90	0.80	0.29 to 1.00	90	0.78	0.29 to 1.00
B2 vs. C2	80	0.67	0.32 to 1.00	90	0.80	0.55 to 1.00
B2 vs. D2	80	0.67	0.31 to 1.00	90	0.80	0.55 to 1.00
B2 vs. E2	80	0.64	0.17 to 1.00	80	0.58	0.00 to 1.00
C2 vs. D2	60	0.40	0.09 to 0.85	80	0.60	0.17 to 1.00
C2 vs. E2	70	0.52	0.23 to 1.00	90	0.80	0.60 to 1.00
D2 vs. E2	70	0.47	−0.13 to 1.00	70	0.40	−0.07 to 1.00
Intra-observer						
A1 vs. A2	60	0.11	−0.31 to 0.74	60	0.05	−0.43 to 0.74
B1 vs. B2	100	1.00		100	1.00	
C1 vs. C2	90	0.83	0.51 to 1.00	90	0.80	0.60 to 1.00
D1 vs. D2	100	1.00		100	1.00	
E1 vs. E2	90	0.82	0.47 to 1.00	100	1.00	

A1 = observer A at rating time-point 1, B2 = observer B at rating time-point 2, etc.
The nominal response categories were N (normal age-specific GMs) AF (abnormal fidgety movements), F− (absence of fidgety movements).
The dichotomous rating was normal/abnormal.

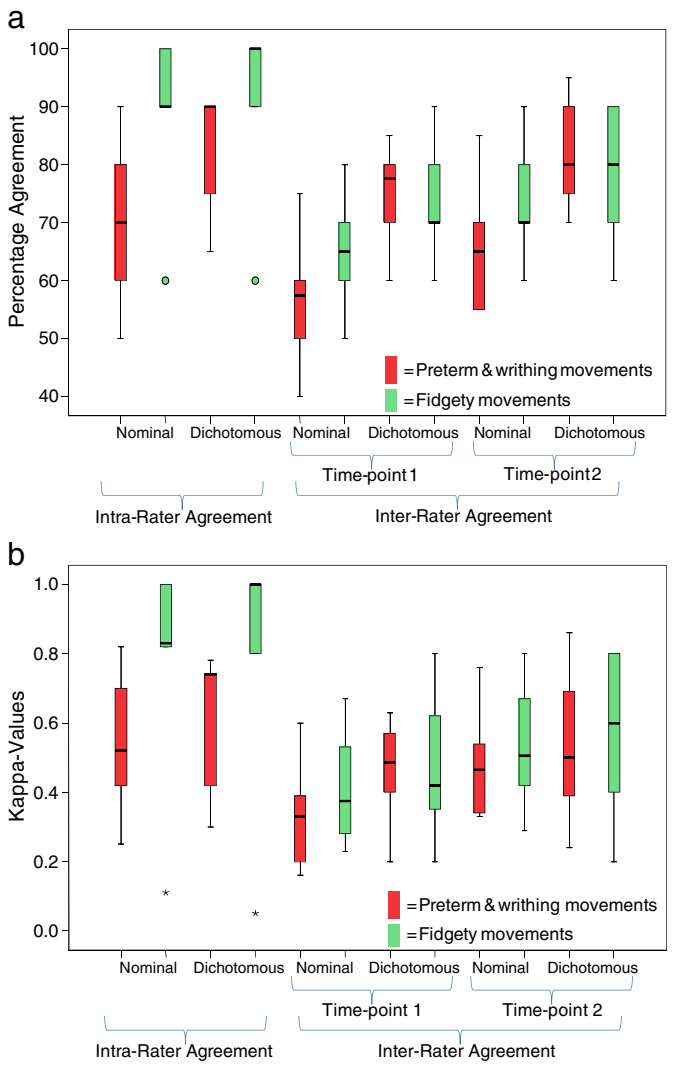

Fig. 2. (a) Boxplots showing percentage agreements. Boxes consist of the five intra-observer agreements and the ten different comparisons of the pairs of observers (inter-observer agreement) for the nominal (for preterm and writhing movements: N (normal age-specific GMs), PR (poor repertoire), Ch (chaotic); CS (cramped synchronized), for fidgety movements: N (normal age-specific GMs), AF (abnormal fidgety movements), F— (absence of fidgety movements), and dichotomous (normal/abnormal) ratings. Red boxes = pairs of observers of the writhing movements, green = fidgety movements. For the intra-observer agreement, there were five pairs (first four boxes). For the inter-observer agreements, there were 10 pairs in each box. Rating time-point 2 was 9 weeks after rating time-point 1. The bold line in the box indicates the median (50th percentile), the lower boundary of the box indicates the 25th percentile and the upper limit of the box the 75th percentile. The box indicates the interquartile range (IQR). The whiskers indicate 1.5 times the IQR, i.e., the lowest or highest value still within 1.5 IQR. The points are extreme values, i.e., between 1.5 and 3 IQR away from the boxes. The stars indicate outliers, more than 3 IQR away from the boxes. (b) Boxplots showing agreements in kappa values. Boxes consist of the five intra-observer agreements and the ten different comparisons of the pairs of observers (inter-observer agreement) for the nominal (for preterm and writhing movements: N (normal age-specific GMs), PR (poor repertoire), Ch (chaotic); CS (cramped synchronized), for fidgety movements: N (normal age-specific GMs), AF (abnormal fidgety movements), F— (absence of fidgety movements), and dichotomous (normal/abnormal) ratings. Red boxes = pairs of observers of the writhing movements, green = fidgety movements. For the intra-observer agreement, there were five pairs (first four boxes). For the inter-observer agreements, there were 10 pairs in each box. Rating time-point 2 was 9 weeks after rating time-point 1. The bold line in the box indicates the median (50th percentile), the lower boundary of the box indicates the 25th percentile and the upper limit of the box the 75th percentile. The box indicates the interquartile range (IQR). The whiskers indicate 1.5 times the IQR, i.e., the lowest or highest value still within 1.5 IQR. The points are extreme values, i.e., between 1.5 and 3 IQR away from the boxes. The stars indicate outliers, more than 3 IQR away from the boxes.

4. Discussion

In this reliability study of the general movements, we obtained the following results: (a) inter-observer agreement was fair to substantial for the six response category ratings and fair to almost perfect for the normal/abnormal ratings. That is to say that the median kappa was below our definition of an acceptable value (substantial, kappa>0.6), except for the normal/abnormal ratings at time-point 2. There was statistically significant improvement from rating time-point 1 to rating time-point 2 for the six response categories. It is unclear why, because we did the same preparation both times, there was no other coordination, discussion or standardization within the team. Perhaps it may be a learning effect based on intensive evaluation in the daily routine. (b) Moderate to almost perfect intra-observer reliability was found for the 9-week interval. That is to say that the median intra-observer agreement was above the acceptable value (substantial, kappa>0.6). (c) The agreements were higher for the normal/abnormal ratings compared to the six response category ratings but not significantly. In order to predict the subsequent neurological outcome of an infant it would be ideal to differentiate the 6 categories. But this is to say that at the preterm and writhing movements age it is already good to differentiate the normals from abnormals because it is always based on developmental trajectories rather than a single recording [16,19]. At the fidgety movements age it is a different situation because it is one recording and normal FM are highly predictive of a normal neurological outcome independent of the GM quality assessed during preterm and writhing movements period [16]. One reason why inter- and intra-observer reliability does not correspond perfectly could be that the video sequences were not always typical examples of the categories but hybrid forms. This could leave space for interpretation. However, this corresponds to everyday life in the clinic. In the training courses and learning videos the excerpts shown are mostly typical and distinct examples. It was found that intra-observer reliability was better the longer the observer had participated in the basic course and the more experienced he was. Even though testers D and E showed almost perfect intra-observer reliability, their inter-observer reliability was only moderate. The inter-observer results between the two observers who had attended the advanced training course were not better than the overall results (Table 3). Many previous studies showing an almost perfect inter-observer agreement (kappa values>0.8 or>90%) and intra-observer agreement (kappa values 0.85–1) were done by members of the GM Trust [6–10,12–15,24]. They are very specialized and experienced professionals who are also used to working scientifically with this assessment. In another study, which shows a fair to almost perfect inter-observer agreement (kappa values 0.36–0.84), the videos were not evaluated by members of the GM Trust and evaluation was focused on not only global but also various other items [23]. Another study with high inter-observer agreement (kappa values 0.78 or 83%) was realized by analyzing the results from the examinations of the basic courses [12]. During the 4 days of training, the main topic is to evaluate videos, and as mentioned before, the videos shown contain typical examples of the different categories. Sometimes the video sequences of the examinations had already been shown and discussed during the course. We asked ourselves whether the different approaches to evaluation might result from the fact that the participants had all been trained by instructors of the GM Trust but not all by the same person.

In our study, the agreement was better in the fidgety movements compared to the preterm and writhing movements groups. Because we rated only 10 videos with fidgety movements, this result has to be regarded with caution. Further research should analyze the difference in agreement between the different GMs based on an adequate number of videos.

5. Conclusion

We were not able to reproduce the generally very good results reported for Prechtl's method on the qualitative assessment of general

movements in preterm, term and young infants. Even so we are convinced that this global assessment can be used as a practicable and helpful clinical tool in combination with physical and neurological examination [29]. In order to continue using GMs as a useful, time-saving, low-cost method of identifying infants for early intervention, in our clinical setting now videos are evaluated by at least two trained therapists and, if necessary, the results are opened up to discussion so that a consensus can be reached. Once a month we get together to watch educational DVD and to discuss the difficult ratings we did in the clinical daily routine. With this procedure, we hope that our evaluation methods will become more and more reliable in the future. Admittedly this does not automatically mean that our results are always correct. The better reliability and validity of Prechtl's method, the earlier the detection of motor impairments and therewith an earlier start of therapy. Further on, there should also be shown atypical and difficult video examples during the training courses. Likewise the evaluation criteria should be more standardized and objectified, although this may prove very difficult because focussing on too many details would disturb the Gestalt perception, which is the background of Prechtl's method [30]. The challenge for the future will be to find an alternative method which is more objective and reliable. At the Bern University Hospital it is planned to run a study which will analyze the correlation between the spontaneous movement of term and preterm infants and neurological outcome. Movements will be measured by three-dimensional accelerometers in this upcoming study.

Conflict of interest statement

All authors hereby declare that they have no financial and personal relationships with other people or organizations that could inappropriately influence (bias) this work.

Acknowledgments

We would like to thank Madeleine Wolf for preparing the video sequences; Isabelle Fankhauser, Marie-Paula Feller, Beatrice Ziswiler for ratings and helpful discussions; and Christina Schläppi and Joy Buchanan for editing the manuscript.

References

[1] Bhutta AT, Cleves MA, Casey PH, Cradock MM, Anand KJ. Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. JAMA Aug 14 2002;288(6):728–37.
[2] Vohr BR, Wright LL, Poole WK, McDonald SA. Neurodevelopmental outcomes of extremely low birth weight infants <32 weeks' gestation between 1993 and 1998. Pediatrics Sep 2005;116(3):635–43.
[3] Spittle AJ, Orton J, Doyle LW, Boyd R. Early developmental intervention programs post hospital discharge to prevent motor and cognitive impairments in preterm infants. Cochrane Database Syst Rev 2007;2(CD005495).
[4] Kanda T, Pidcock FS, Hayakawa K, Yamori Y, Shikata Y. Motor outcome differences between two groups of children with spastic diplegia who received different intensities of early onset physiotherapy followed for 5 years. Brain Dev Mar 2004;26(2):118–26.
[5] Blauw-Hospers CH, de Graaf-Peters VB, Dirks T, Bos AF, Hadders-Algra M. Does early intervention in infants at high risk for a developmental motor disorder improve motor and cognitive development? Neurosci Biobehav Rev 2007;31(8):1201–12.
[6] Cioni G, Prechtl HF, Ferrari F, Paolicelli PB, Einspieler C, Roversi MF. Which better predicts later outcome in full-term infants: quality of general movements or neurological examination? Early Hum Dev Nov 24 1997;50(1):71–85.
[7] Einspieler C, Cioni G, Paolicelli PB, Bos AF, Dressler A, Ferrari F, et al. The early markers for later dyskinetic cerebral palsy are different from those for spastic cerebral palsy. Neuropediatrics Apr 2002;33(2):73–8.
[8] Bos AF, van Asperen RM, de Leeuw DM, Prechtl HF. The influence of septicaemia on spontaneous motility in preterm infants. Early Hum Dev Nov 24 1997;50(1):61–70.
[9] Bos AF, van Loon AJ, Hadders-Algra M, Martijn A, Okken A, Prechtl HF. Spontaneous motility in preterm, small-for-gestational age infants. II. Qualitative aspects. Early Hum Dev Nov 24 1997;50(1):131–47.
[10] Bos AF, Martijn A, Okken A, Prechtl HF. Quality of general movements in preterm infants with transient periventricular echodensities. Acta Paediatr Mar 1998;87(3):328–35.
[11] Burger M, Louw QA. The predictive validity of general movements—a systematic review. Eur J Paediatr Neurol Sep 2009;13(5):408–20.
[12] Valentin T, Uhl K, Einspieler C. The effectiveness of training in Prechtl's method on the qualitative assessment of general movements. Early Hum Dev Jul 2005;81(7):623–7.
[13] Einspieler C, Prechtl HF, Ferrari F, Cioni G, Bos AF. The qualitative assessment of general movements in preterm, term and young infants—review of the methodology. Early Hum Dev Nov 24 1997;50(1):47–60.
[14] Fjortoft T, Einspieler C, Adde L, Strand LI. Inter-observer reliability of the "Assessment of Motor Repertoire—3 to 5 Months" based on video recordings of infants. Early Hum Dev May 2009;85(5):297–302.
[15] Spittle AJ, Doyle LW, Boyd RN. A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life. Dev Med Child Neurol Apr 2008;50(4):254–66.
[16] Einspieler C, Prechtl HF, Bos AF, Ferrari F, Cioni G. Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants. Mac Keith Press; 2004.
[17] Prechtl HF. State of the art of a new functional assessment of the young nervous system. An early predictor of cerebral palsy. Early Hum Dev Nov 24 1997;50(1):1–11.
[18] Prechtl HF. General movement assessment as a method of developmental neurology: new paradigms and their consequences. The 1999 Ronnie MacKeith lecture. Dev Med Child Neurol Dec 2001;43(12):836–42.
[19] Prechtl HF. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. Early Hum Dev Sep 1990;23(3):151–8.
[20] Lorenz K. Gestalt perception as a source of scientific knowledge. In: Lorenz K, editor. Studies in Animal and Human Behaviour. London: Methuen; 1971. p. 281–322.
[21] Ferrari F, Cioni G, Einspieler C, Roversi MF, Bos AF, Paolicelli PB, et al. Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy. Arch Pediatr Adolesc Med May 2002;156(5):460–7.
[22] Romeo DM, Guzzetta A, Scoto M, Cioni M, Patusi P, Mazzone D, et al. Early neurologic assessment in preterm-infants: integration of traditional neurologic examination and observation of general movements. Eur J Paediatr Neurol May 2008;12(3):183–9.
[23] van Kranen-Mastenbroek V, van Oostenbrugge R, Palmans L, Stevens A, Kingma H, Blanco C, et al. Inter- and intra-observer agreement in the assessment of the quality of spontaneous movements in the newborn. Brain Dev Sep 1992;14(5):289–93.
[24] Mutlu A, Einspieler C, Marschik P, Livanelioglu A. Intra-individual consistency in the quality of neonatal general movements. Neonatology 2007;93(3):213–6.
[25] Hadders-Algra M. General movements: a window for early identification of children at high risk for developmental disorders. J Pediatr Aug 2004;145(Suppl 2):S12–8.
[26] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics Mar 1977;33(1):159–74.
[27] Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther Mar 2005;85(3):257–68.
[28] Reichenheim ME. sxd3: Sample size for the kappa-statistic of interrater agreement. Stata Tech Bull 2000;s8:41–5.
[29] Heineman KR, Hadders-Algra M. Evaluation of neuromotor function in infancy—a systematic review of available methods. J Dev Behav Pediatr Aug 2008;29(4):315–23.
[30] Nakajima Y, Einspieler C, Marschik PB, Bos AF, Prechtl HF. Does a detailed assessment of poor repertoire general movements help to identify those infants who will develop normally? Early Hum Dev Jan 2006;82(1):53–9.