Acta Pædiatrica, 2006; Suppl 450: 47–55

Taylor & Francis
Taylor & Francis Group

Reliability of motor development data in the WHO Multicentre Growth Reference Study

WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP[1,2]

[1]*Department of Nutrition, World Health Organization, Geneva, Switzerland, and,* [2]*Members of the WHO Multicentre Growth Reference Study Group (listed at the end of the first paper in this supplement)*

Abstract

Aim: To describe the methods used to standardize the assessment of motor milestones in the WHO Multicentre Growth Reference Study (MGRS) and to present estimates of the reliability of the assessments. *Methods:* As part of the MGRS, longitudinal data were collected on the acquisition of six motor milestones by children aged 4 to 24 mo in Ghana, India, Norway, Oman and the USA. To ensure standardized data collection, the sites conducted regular standardization sessions during which fieldworkers took turns to examine and score about 10 children for the six milestones. Assessments of the children were videotaped, and later the other fieldworkers in the same site watched the videotaped sessions and independently rated performances. The assessments were also viewed and rated by the study coordinator. The coordinator's ratings were considered the reference (true) scores. In addition, one cross-site standardization exercise took place using videotapes of 288 motor assessments. The degree of concordance between fieldworkers and the coordinator was analysed using the Kappa coefficient and the percentage of agreement. *Results:* Overall, high percentages of agreement (81–100%) between fieldworkers and the coordinator and "substantial" (0.61–0.80) to "almost perfect" (>0.80) Kappa coefficients were obtained for all fieldworkers, milestones and sites. Homogeneity tests confirm that the Kappas are homogeneous across sites, across milestones, and across fieldworkers. Concordance was slightly higher in the cross-site session than in the site standardization sessions. There were no systematic differences in assessing children by direct examination or through videotapes.

Conclusion: These results show that the criteria used to define performance of the milestones were similar and applied with equally high levels of reliability among fieldworkers within a site, among milestones within a site, and among sites across milestones.

Key Words: *Agreement, children, inter-rater reliability, motor development, motor skills*

Introduction

The World Health Organization (WHO), in collaboration with partner institutions worldwide, conducted the WHO Multicentre Growth Reference Study (MGRS) to generate new growth curves for assessing the growth and development of infants and young children [1]. As part of the longitudinal component of the MGRS, the Motor Development Study (MDS) was carried out to assess the acquisition of six distinct key motor milestones by affluent children growing up in different cultures. The assessments were done from 4 mo of age until the children were able to walk independently, or reached 24 mo, in Ghana, India, Norway, Oman and the USA. The details of the MDS's study design and methodology have been described elsewhere [2]. To our knowledge, only two other multi-country studies of motor development have used a longitudinal design [3,4].

Rigorous data collection procedures and quality-control measures were applied in all sites to minimize measurement error when assessing motor milestone achievement and to avoid bias among sites. Variability in methods of measurement can occur for several reasons [5–7]:

1. *The setting in which the assessments are carried out.* Data collection took place at the children's homes and thus the assessment environment was somewhat variable except for what we could control. Where possible, the number of persons present during assessments was limited to three (fieldworker, caretaker and child); also, the surface of the floor where the assessments took place was kept clean and free of objects that

Correspondence: Mercedes de Onis, Study Coordinator, Department of Nutrition, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. Tel: +41 22 791 3320. Fax: +41 22 791 4156. E-mail: deonism@who.int

ISSN 0803-5326 print/ISSN 1651-2227 online
DOI: 10.1080/08035320500495480

might interfere with locomotion, and a maximum of three toys or objects with which the child liked to play were available [2].

2. *The child's mood.* Children vary in their emotional state during assessments for a variety of reasons, and this cannot be controlled. Care was taken, however, to reassure and calm the children and to record their overall emotional state according to two scales described by Brazelton [8].
3. *The examiner's mood.* Examiners also vary among themselves, and over time, in mood, level of energy and motivation. Efforts were made to keep fieldworkers motivated, to impress upon them the importance of the study, and to repeatedly emphasize the need to adhere to the standardized protocol. In addition, appropriate training, site visits by the MDS coordinator and monitoring of data quality were essential to control for this third possible source of variability and to minimize bias across sites.
4. *Methodological differences among fieldworkers.* Observational assessment tools such as the assessment of motor milestones are particularly prone to error due to differences among fieldworkers in judging when a particular behaviour has been exhibited [9]. Therefore, considerable effort was made to standardize the criteria for assessing when certain motor skills were demonstrated, such as clear instructions and drawings in the procedures manual, periodic standardization sessions in all sites, and the use of videotapes to standardize criteria across sites.

The purpose of this paper is to describe the methods used to standardize the assessment of motor milestones in the MGRS and to present estimates of the reliability of these assessments.

Methods

Periodic site standardization sessions

Standardization sessions were conducted on a regular basis (at 1-mo or 2-mo intervals) during data collection in Ghana, India, Norway and Oman. The North American site did so only once because data collection was nearly completed by the time the decision was taken to conduct regular standardization sessions; also, and for the same reason, this site did not participate in the cross-site standardization exercise. Due to limited data availability, the North American site was thus not included in the analyses for this paper. Brazil, which was the earliest MGRS site, did not assess motor milestones.

During each session, 10 apparently healthy children, aged 6 to 12 mo, were recruited for participation through day-care and health centres. At every session, one of the fieldworkers examined and scored the children for each of the six gross motor milestones: sitting without support, hands-and-knees crawling, standing with assistance, walking with assistance, standing alone and walking alone. A different fieldworker was selected for each session to give everyone a turn. The performance of each milestone was recorded as follows: "inability"—the child tried but failed to perform the test item; "ability"—the child performed the test item according to the specified criteria; "refusal"—the child was calm and alert but uncooperative; and "unable to test"—the child could not be examined because his or her emotional state (drowsiness, fussiness or crying) interfered with the examination or the child's caretaker was distraught. In practice, it proved difficult to distinguish between "refusal" and "unable to test", and these categories were therefore combined. The child's caregiver was present during all assessments but was requested not to interfere with the examination. However, when needed, the examiner asked for the caregiver's assistance, for instance in placing the child into the correct position or in encouraging the child to crawl or walk. The examiner recorded the results discretely, taking care not to disclose the child's rating. Since it was not always possible to get the child to cooperate immediately, the examiner was allowed three tries to assess each milestone.

Assessments of the children were videotaped, and later the other fieldworkers in the same site watched the videotaped sessions and independently rated performances. The videotape of the session and the fieldworkers' ratings were then sent to the MGRS Coordinating Centre at WHO in Geneva where the MDS coordinator viewed the tape and rated the children's performance. The ratings given by the coordinator were considered to be the reference (true) scores.

Cross-site standardization session

The MDS coordinator visited Ghana, India, Norway and Oman to carry out standardization exercises using videotapes of 288 motor assessments made in 51 children. Care was taken to select the best demonstrations of the milestones. The fieldworkers in all four countries viewed the videotapes and independently rated the children's performance.

Statistical analysis

Three outcome categories were examined: 1) observed inability; 2) refusal and/or unable to test; and 3) observed ability.

The degree of concordance between fieldworkers and the MDS coordinator was analysed using the

Kappa (κ) coefficient, a measure of association for categorical variables [10]. Kappa compares the observed agreement between pairs of raters to the agreement expected by chance when judgements are statistically independent [11]. Kappa coefficients vary between 0 and 1. A Kappa coefficient of ≤ 0.20 indicates slight agreement, $\kappa = 0.21 - 0.40$ indicates fair agreement, $\kappa = 0.41 - 0.60$ indicates moderate agreement, $\kappa = 0.61 - 0.80$ indicates substantial agreement and $\kappa > 0.80$ means almost perfect agreement [12].

The percentage of agreement was also estimated because this value can be calculated in all instances [13], whereas Kappa coefficients cannot be calculated if all children are rated similarly by both fieldworkers. The percentage of agreement was calculated by dividing the number of agreements between a fieldworker's rating and the MDS coordinator by the total number of paired observations [13]. Agreement of 90% or more was considered high [2].

Further analysis was based on the methodology suggested by Reed [14] that allows one to judge whether the Kappa coefficients from several studies or clinical centres "belong together" as a set. In the MDS, a key question is whether Kappa coefficients across participating sites pass the homogeneity test. The null hypothesis is that the Kappas of all sites are equal for each of the milestones (H_0: $\kappa_{Ghana} = \kappa_{India} = \kappa_{Norway} = \kappa_{Oman}$). For this purpose, summary Kappa coefficients were calculated for all fieldworkers within a site and for each milestone. The goodness-of-fit test of the null hypothesis H_0 was obtained by using a statistic that is assumed to be χ^2 distributed with n (= number of sites -1) degrees of freedom. Homogeneity was also assessed for Kappa coefficients across fieldworkers within sites and for each milestone (i.e. do all fieldworkers within a site have similar Kappas for each milestone?) and across milestones within sites (i.e. are the Kappas similar within sites for all six milestones?).

Two sources of information are available about concordance in the ratings of motor milestones between fieldworkers and the MDS coordinator: the site-specific exercises and the cross-site session. Should similar Kappa coefficients be expected? To answer this question, differences in approaches must be considered. All assessments by all fieldworkers in all sites used the same set of videotapes in the cross-site standardization session, whereas the site standardization sessions included local children and assessments by fieldworkers were done either by direct examination of the child or through videotapes. The MDS coordinator assessed video recordings in both types of exercises, although she was present in the sites during the cross-site standardization session. Because the videos were selected for teaching purposes, including clarity in filming and in the demonstration of motor behaviours, better concordance between fieldworkers and the MDS coordinator might be expected in the cross-site session.

Finally, we examined the level of concordance with the MDS coordinator in the rating of motor milestones when fieldworkers assessed children by direct examination or through videotapes by randomly selecting three fieldworkers per site and comparing their Kappa coefficients and percentage of agreement in each site.

All statistical analyses were performed using Stata 8.0 [15].

Results

Periodic site standardization sessions

Kappa coefficients and percent agreement with the MDS coordinator are given in Table I for all fieldworkers, by site, across all standardization sessions. The number of sessions varied by site: Ghana 8, India 11, Norway 2 and Oman 11. The number of children assessed per fieldworker and milestone varied as well because some fieldworkers did not complete the standardization sessions or because some milestone assessments were omitted due to poor filming. In general, there were "substantial" to "almost perfect" levels of agreement between fieldworkers and the MDS coordinator across all milestones and sites. Exceptions were the Kappa coefficients for the milestone "sitting without support" for fieldworker no. 4 in Ghana ($\kappa = 0.585$) and for the milestones "standing alone" and "walking alone" for fieldworker no. 6 in Norway ($\kappa = 0.422$ and 0.345, respectively). The percentage of agreement ranged between 81.0% (Norway, standing with assistance) and 100.0%.

Cross-site standardization session

Table II presents similar data to that in Table I but for the cross-site standardization session, where the MDS coordinator travelled to the sites and showed the same videotapes of 288 motor assessments. The Kappa coefficients indicate "substantial" to "almost perfect" levels of agreement between fieldworkers and the MDS coordinator. The percentage of agreement ranged between 80.9% (Ghana, walking alone) and 100.0%.

Concordance was rated "substantial" to "almost perfect" in both the periodic site and the cross-site standardization sessions but was often slightly higher in the cross-site session for all milestones except "walking alone" (values in Table II tend to be greater than values in Table I).

Table I. Kappa coefficients and % of agreement with the MDS coordinator for all fieldworkers, by site, for the periodic site standardization sessions[a].

		Ghana			India			Norway			Oman		
	Fieldworker	*n*	Kappa	% agree	*n*	Kappa	% agree	*n*	Kappa	% agree	*n*	Kappa	% agree
Sitting without support	1	83	0.851	98.8	107	0.904	98.1	20	0.857	95.0	103	0.923	97.1
	2	63	1.000	100.0	39	0.898	97.4	20	0.857	95.0	103	0.949	98.1
	3	53	0.660	98.1	107	0.900	98.1	20	0.857	95.0	103	0.925	97.1
	4	83	0.585	95.2	107	1.000	100.0	20	0.771	90.0	103	0.950	98.1
	5	53	0.658	98.1	107	0.952	99.1	20	1.000	100.0	103	1.000	100.0
	6	83	0.851	98.8	107	0.952	99.1	20	0.857	95.0	103	0.951	98.1
	7	83	0.851	98.8	39	0.898	97.4	20	1.000	100.0			
	8				77	0.892	97.4						
	9				107	0.908	98.1						
	Overall	501	0.761	98.2	797	0.927	98.5	140	0.884	95.7	618	0.950	98.1
Hands-and-knees crawling	1	84	0.960	97.6	105	0.949	97.1	22	0.919	95.5	106	0.939	96.2
	2	65	0.949	96.9	35	0.849	91.4	22	0.919	95.5	106	0.939	96.2
	3	55	0.820	89.1	105	0.880	93.3	22	0.833	90.9	106	0.970	98.1
	4	84	0.800	88.1	105	0.966	98.1	22	1.000	100.0	106	0.954	97.2
	5	55	0.938	96.4	105	0.931	96.2	22	1.000	100.0	106	0.955	97.2
	6	84	0.960	97.6	105	0.883	93.3	22	0.919	95.5	106	0.924	95.3
	7	84	0.942	96.4	35	0.952	97.1	22	1.000	100.0			
	8				75	0.861	92.0						
	9				105	0.897	94.3						
	Overall	511	0.912	94.7	775	0.911	95.0	154	0.943	96.8	636	0.947	96.7
Standing with assistance	1	74	0.808	89.2	97	0.830	91.8	21	0.837	90.5	100	0.857	92.0
	2	57	0.727	86.0	38	0.867	92.1	21	1.000	100.0	100	0.875	93.0
	3	48	0.826	91.7	97	0.831	91.8	21	0.837	90.5	100	0.893	94.0
	4	74	0.738	86.5	97	0.809	90.7	21	0.837	90.5	100	0.892	94.0
	5	48	0.767	87.5	97	0.785	89.7	21	1.000	100.0	100	0.911	95.0
	6	74	0.813	90.5	97	0.894	94.8	21	0.653	81.0	100	0.892	94.0
	7	74	0.760	87.8	38	0.869	92.1	21	0.755	85.7			
	8				71	0.893	94.4						
	9				97	0.804	90.7						
	Overall	449	0.777	88.4	729	0.839	91.9	147	0.847	91.2	600	0.887	93.7
Walking with assistance	1	76	0.905	94.7	104	0.792	87.5	20	1.000	100.0	104	0.793	86.5
	2	60	0.822	90.0	37	0.903	94.6	20	0.917	95.0	104	0.777	85.6
	3	50	0.891	94.0	104	0.839	90.4	20	1.000	100.0	104	0.912	94.2
	4	76	0.902	94.7	104	0.869	92.3	20	0.917	95.0	104	0.729	82.7
	5	50	0.854	92.0	104	0.836	90.4	20	1.000	100.0	104	0.807	87.5
	6	76	0.856	92.1	104	0.889	93.3	20	0.817	90.0	104	0.808	87.5
	7	76	0.882	93.4	37	0.808	89.2	20	1.000	100.0			
	8				74	0.841	90.5						
	9				104	0.773	86.5						
	Overall	464	0.875	93.1	772	0.838	90.3	140	0.951	97.1	624	0.805	87.3

Table I (*Continued*)

		Ghana			India			Norway			Oman		
	Fieldworker	*n*	Kappa	% agree	*n*	Kappa	% agree	*n*	Kappa	% agree	*n*	Kappa	% agree
Standing alone	1	72	0.926	95.8	108	0.736	86.1	20	1.000	100.0	105	0.919	95.2
	2	57	0.836	91.2	39	0.875	94.9	20	0.683	85.0	105	0.902	94.3
	3	47	0.800	89.4	108	0.845	91.7	20	0.897	95.0	105	0.968	98.1
	4	72	0.897	94.4	108	0.863	92.6	20	0.797	90.0	105	0.798	88.6
	5	47	0.783	89.4	108	0.768	88.0	20	0.797	90.0	105	0.902	94.3
	6	72	0.850	91.7	108	0.884	93.5	20	0.422	75.0	105	0.936	96.2
	7	72	0.873	93.1	39	0.939	97.4	20	0.785	90.0			
	8				78	0.725	87.2						
	9				108	0.823	90.7						
	Overall	439	0.861	92.5	804	0.820	90.7	140	0.776	89.3	630	0.905	94.4
Walking alone	1	60	0.902	95.0	109	0.732	88.1	19	0.835	94.7	106	0.851	91.5
	2	45	0.773	88.9	40	0.804	92.5	19	0.835	94.7	106	0.834	90.6
	3	35	0.743	88.6	109	0.838	92.7	19	0.835	94.7	106	0.950	97.2
	4	60	0.867	93.3	109	0.895	95.4	19	0.835	94.7	106	0.741	85.8
	5	35	0.743	88.6	109	0.820	92.7	19	1.000	100.0	106	0.896	94.3
	6	72	0.827	91.7	109	0.867	93.6	19	0.345	84.2	106	0.821	89.6
	7	72	0.880	94.4	40	0.939	97.5	19	1.000	100.0			
	8				79	0.806	92.4						
	9				109	0.849	93.6						
	Overall	379	0.835	92.1	813	0.835	92.9	133	0.822	94.7	636	0.849	91.5

[a] Analyses combine all standardization sessions per site (8 in Ghana, 11 in India, 2 in Norway and 11 in Oman).

Table II. Kappa coefficients and % of agreement with the MDS coordinator for all fieldworkers, by site, for the cross-site standardization session using videotapes of 288 motor assessments.

		Ghana		India		Norway		Oman	
	Fieldworker	Kappa	% agree	Kappa	%agree	Kappa	% agree	Kappa	% agree
Sitting without support ($n = 49$)	1	1.000	100.0	1.000	100.0	0.866	95.9	1.000	100.0
	2	0.930	98.0	0.936	98.0	0.930	98.0	1.000	100.0
	3	0.930	98.0	0.826	93.9	0.867	95.9	1.000	100.0
	4	0.871	95.9	0.936	98.0	0.879	95.9	0.930	98.0
	5	0.854	95.9	0.936	98.0	0.877	95.9	0.657	87.8
	6	1.000	100.0	1.000	100.0	0.936	98.0		
	7	0.868	95.9	1.000	100.0	0.867	95.9		
	8			1.000	100.0				
	Overall	0.923	97.7	0.952	98.5	0.889	96.5	0.909	97.1
Hands-and-knees crawling ($n = 47$)	1	0.894	93.6	0.964	97.9	0.887	93.6	0.887	93.6
	2	1.000	100.0	0.963	97.9	0.887	93.6	0.887	93.6
	3	0.893	93.6	0.964	97.9	0.735	85.1	0.926	95.7
	4	0.812	89.4	0.927	95.7	0.928	95.7	0.926	95.7
	5	0.963	97.9	0.859	91.5	0.926	95.7	0.776	87.2
	6	0.963	97.9	0.891	93.6	0.852	91.5		
	7	0.928	95.7	0.890	93.6	0.854	91.5		
	8			0.964	97.9				
	Overall	0.922	95.4	0.924	95.5	0.867	92.4	0.880	93.2
Standing with assistance ($n = 51$)	1	0.837	90.2	0.896	94.1	0.746	86.3	0.931	96.1
	2	0.864	92.2	0.828	90.2	0.896	94.1	0.860	92.2
	3	0.896	94.1	0.932	96.1	0.859	92.2	0.896	94.1
	4	0.827	90.2	0.824	90.2	0.863	92.2	0.895	94.1
	5	0.901	94.1	0.863	92.2	0.899	94.1	0.861	92.2
	6	0.897	94.1	0.933	96.1	0.720	84.3		
	7	0.862	92.2	0.898	94.1	0.862	92.2		
	8			0.896	94.1				
	Overall	0.869	92.4	0.888	93.6	0.836	90.8	0.889	93.7
Walking with assistance ($n = 48$)	1	0.962	97.9	0.818	89.6	0.889	93.8	1.000	100.0
	2	0.927	95.8	0.814	89.6	0.890	93.8	0.925	95.8
	3	0.924	95.8	0.890	93.8	0.769	87.5	0.963	97.9
	4	0.962	97.9	0.887	93.8	0.852	91.7	0.887	93.8
	5	0.888	93.8	0.846	91.7	0.887	93.8	0.962	97.9
	6	0.962	97.9	0.925	95.8	0.888	93.8		
	7	0.925	95.8	0.890	93.8	0.927	95.8		
	8			0.753	85.4				
	Overall	0.935	96.4	0.848	91.4	0.872	92.9	0.947	97.1
Standing alone ($n = 46$)	1	0.952	97.8	0.901	95.7	0.819	91.3	1.000	100.0
	2	0.902	95.7	1.000	100.0	0.949	97.8	0.901	95.7
	3	0.857	93.5	0.907	95.7	0.896	95.7	0.951	97.8
	4	0.648	84.8	1.000	100.0	1.000	100.0	0.952	97.8
	5	0.949	97.8	0.952	97.8	0.902	95.7	0.851	93.5
	6	0.949	97.8	0.952	97.8	0.848	93.5		
	7	0.951	97.8	1.000	100.0	0.763	89.1		
	8			0.951	97.8				
	Overall	0.888	95.0	0.964	98.4	0.881	94.7	0.931	97.0
Walking alone ($n = 47$)	1	0.801	93.6	0.678	89.4	0.702	89.4	0.803	93.6
	2	0.780	93.6	0.931	97.9	0.927	97.9	0.803	93.6
	3	0.721	91.5	0.702	89.4	0.780	93.6	0.861	95.7
	4	0.722	91.5	1.000	100.0	0.927	97.9	0.801	93.6
	5	0.780	93.6	1.000	100.0	0.794	93.6	0.813	93.6
	6	0.780	93.6	0.771	93.6	0.781	93.6		
	7	0.861	95.7	0.862	95.7	0.658	87.2		
	8			0.861	95.7				
	Overall	0.778	93.3	0.838	95.0	0.788	93.3	0.816	94.0

Table III. Tests of homogeneity of Kappa coefficients in the MDS: *p*-values for the periodic site standardization sessions (SSS) and for the cross-site standardization session (CSS).

	Ghana		India		Norway		Oman		Across sites, within milestones
	SSS	CSS	SSS	CSS	SSS	CSS	SSS	CSS	CSS
Sitting without support	NA[a]	0.619	0.925	0.246	0.789	0.848	0.580	NA[b]	0.414
Hands-and-knees crawling	0.198	0.265	0.497	0.646	0.602	0.550	0.903	0.477	0.274
Standing with assistance	0.942	0.983	0.926	0.900	0.355	0.510	0.989	0.916	0.463
Walking with assistance	0.923	0.912	0.772	0.665	0.420	0.790	0.519	0.418	0.082
Standing alone	0.857	0.050	0.613	0.629	0.619	0.318	0.127	0.501	0.084
Walking alone	0.753	0.305	0.656	0.102	0.768	0.452	0.116	0.955	0.890
Across milestones, within sites	0.199	0.546	0.438	0.668	0.384	0.772	0.265	0.662	

[a] Test of homogeneity among Kappas can not be performed because the number of concordant negative ratings (i.e. fieldworker and MDS coordinator recording that the child was unable to perform the milestone) was zero for all fieldworkers for milestone sitting without support.
[b] Test of homogeneity among Kappas can not be performed because the number of discordant (i.e. fieldworker and MDS coordinator recording different ratings for the same child) was zero for three out of five fieldworkers for milestone sitting without support.

Homogeneity

Table III presents results assessing the homogeneity of Kappa coefficients in the site standardization sessions and the cross-site session. *P*-values inside the table (all values but those given in the bottom row and right-hand column) answer the question: Are the fieldworkers homogeneous in assessing motor milestones within a site? *P*-values in the right-hand column answer the question: Are the fieldworkers homogeneous in assessing motor milestones across sites when viewing the same videotapes? *P*-values on the bottom row answer the question: Are the fieldworkers homogeneous in their assessments across milestones within a site? None of the *P*-values were statistically significant ($p < 0.05$), although one value (Ghana, standing alone, CSS) had a *p*-value of 0.05. These results indicate that the Kappas are homogeneous across sites, across milestones, and across fieldworkers.

Concordance in assessment by direct examination versus videotape

Table IV presents, for 12 randomly selected fieldworkers (three per site), the Kappa coefficients and percentage of agreement with the MDS coordinator when fieldworkers tested children by direct examination or using videotapes. Overall, there were no systematic differences to indicate that one way of conducting the assessment is more concordant with the MDS coordinator than the other.

Discussion

This is the first longitudinal study to use a standardized protocol to describe gross motor development among healthy children from different countries and to carry out standardization sessions on a regular basis. Kappa coefficients were used to estimate the concordance of independent pairs of raters, specifically one of several fieldworkers and always the MDS coordinator. These values estimate the quality of the MDS testing procedures [2] and the fieldworkers' ability to apply the rating criteria consistently.

Overall, high percentages of agreement between fieldworkers and the MDS coordinator, and "substantial" to "almost perfect" Kappa coefficients, were obtained for all fieldworkers, milestones and sites. Homogeneity tests confirm that the Kappa coeffi-

Table IV. Comparison of Kappa coefficients and percentage agreement when three randomly selected fieldworkers per site assessed children by direct examination or through videotapes.

Site	Assessment	Milestone[a]	Kappa	% agreement
Ghana	Direct	2	1.000	100.0
	Video		0.945	96.9
Ghana	Direct	2	0.808	90.0
	Video		0.796	87.8
Ghana	Direct	5	0.912	94.1
	Video		0.929	96.4
India	Direct	1	1.000	100.0
	Video		0.948	99.0
India	Direct	2	0.805	87.5
	Video		0.887	93.8
India	Direct	4	0.821	90.0
	Video		0.839	90.4
Norway	Direct	2	1.000	100.0
	Video		0.896	94.1
Norway	Direct	4	1.000	100.0
	Video		0.902	93.8
Norway	Direct	5	0.556	75.0
	Video		0.360	75.0
Oman	Direct	3	0.841	90.0
	Video		0.896	94.4
Oman	Direct	4	0.628	75.0
	Video		0.755	84.5
Oman	Direct	6	0.814	88.9
	Video		0.834	90.9

[a] Milestone: 1 = sitting without support; 2 = hands-and-knees crawling; 3 = standing with assistance; 4 = walking with assistance; 5 = standing alone; 6 = walking alone.

cients are a homogeneous set across sites, across milestones, and across fieldworkers. Concordance was slightly higher in the cross-site session (i.e. when fieldworkers rated the same set of videotapes) than in the periodic site standardization sessions where different sets of local children were assessed.

The forgoing analyses show that the standardization of milestone assessments made in any one site were consistently high among fieldworkers within a site, among milestones within a site, and among sites across all six milestones. Also, the cross-site exercise indicates that the fieldworkers could reliably rate motor milestones of children both in their own and in the other sites.

There are few reports of inter-rater agreement [16–19] in motor milestones assessments, and what information is available suggests that the MDS concordance is very good relative to other studies. For example, the mean percentage of agreement between four examiners during the standardization of the Denver Developmental Screening Test was 90%, with a range of 80–95% [17]. Using the Movement Assessment of Infants, Haley et al. [16] reported only 2% of the items demonstrated excellent ($\kappa > 0.75$) inter-rater reliability beyond chance, with 58% in the fair-to-good ($0.40 < \kappa < 0.75$) range.

The six milestones were selected for the study because they were considered to be both fundamental to the acquisition of self-sufficient erect locomotion and simple to administer and evaluate. They should measure observable behaviour with a clear pass or fail score. The high degree of inter-rater reliability confirms that these milestones were simple to administer and feasible to standardize. These results were probably attributable to the clarity of the instructions for administering and rating the performance of the milestones, and to the fact that fieldworkers were well trained. As observed in other studies [18,19], the multiple standardization sessions no doubt added to the fieldworkers' skills and confidence in conducting motor development assessments.

The organization of reliability sessions is often logistically demanding and places considerable stress on both researchers and family members. An attractive alternative is to estimate inter-rater reliability coefficients with the aid of videotapes instead of having several examiners test a group of children more than once. Stuberg et al. [20] found that minimizing the handling of children and relying on observation help achieve more accurate test results. Children can behave differently from one time to the next [17], and these differences may influence the reliability coefficients. By using videotapes, these results reflected the fieldworker's ability to rate the test items under controlled conditions, that is without having to deal with children's moods and behaviours. On the other hand, Gowland et al. [21] concluded that observing task performances from a videotape appeared to be a major source of variability because taping frequently did not capture the full performance, or part of the body to be observed was not filmed fully or from an appropriate angle. Our study excluded milestone assessments that could not be rated for these reasons, and we found no systematic difference in the Kappa coefficients and percentage of agreement when fieldworkers rated children by direct examination or through videotapes.

We found several advantages, which were also common to other studies [6,22,23], in using video recordings to evaluate rating performances. Videotapes helped to alleviate problems with recruiting children and scheduling sessions. Fieldworkers were able to rate the motor development assessments when convenient to them. The MDS coordinator could examine the tape with the fieldworkers to explore possible reasons for disagreement. Most importantly, children did not have to endure repeated assessments by numerous fieldworkers. Russell et al. [6] cited as a main disadvantage that this method tests only the participant's ability to rate the videotaped assessments but provides no indication of the participant's ability to administer and score them in a clinical or study situation. This is a fair criticism, and for this reason studies should assess the quality of assessments in both direct examination and video settings. This is what we did, but in our case we did not find systematic differences between these settings.

The MDS protocol was designed to provide a simple method of evaluating six gross motor milestones in young children. The WHO MGRS, in implementing this protocol, provided the opportunity to evaluate these milestones in multiple countries and, for the first time, to use the data collected to construct an international standard for the achievement of six universal gross motor development milestones [24,25]. Assessing children's behaviour, including gross motor milestones, is demanding for both fieldworkers and children. The results of this study demonstrate that, with careful attention to protocol and training, a high level of fieldworker reliability can be achieved within and across sites.

Acknowledgements

This paper was prepared by Trudy M.A. Wijnhoven, Mercedes de Onis, Reynaldo Martorell, Edward A. Frongillo and Gunn-Elin A. Bjoerneboe on behalf of the WHO Multicentre Growth Reference Study Group. The statistical analysis was conducted by Amani Siyam.

References

[1] de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martines J, for the WHO Multicentre Growth Reference Study Group. The WHO Multicentre Growth Reference Study: Planning, study design and methodology. Food Nutr Bull 2004;25 Suppl 1:S15–26.

[2] Wijnhoven TM, de Onis M, Onyango AW, Wang T, Bjoerneboe GE, Bhandari N, et al., for the WHO Multicentre Growth Reference Study Group. Assessment of gross motor development in the WHO Multicentre Growth Reference Study. Food Nutr Bull 2004;25 Suppl 1:S37–45.

[3] Hindley CB, Filliozat AM, Klackenberg G, Nicolet-Meister D, Sand EA. Differences in age of walking in five European longitudinal samples. Hum Biol 1966;38:364–79.

[4] World Health Organization, Task Force for Epidemiological Research on Reproductive Health; Special Programme of Research, Development, and Research Training in Human Reproduction. Progestogen-only contraceptives during lactation: II. Infant development. Contraception 1994;50:55–68.

[5] Krebs DE. Measurement theory. Phys Ther 1987;67:1834–9.

[6] Russell DJ, Rosenbaum PL, Lane M, Gowland C, Goldsmith CH, Boyce WF, et al. Training users in the gross motor function measure: methodological and practical issues. Phys Ther 1994;74:630–6.

[7] Plewis A, Bax M. The uses and abuses of reliability measures in developmental medicine. Dev Med Child Neurol 1982;24: 388–90.

[8] Brazelton TB. Echelle d'évaluation du comportement néonatal. Neuropsychiatr Enfance Adolesc 1983;31:61–96.

[9] Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychol Bull 1979;86:376–90.

[10] Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. Stat Med 2002;21:2109–29.

[11] Agresti A. An introduction to categorical data analysis. Wiley series in probability and statistics. New York: John Wiley & Sons, Inc.; 1996.

[12] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

[13] Altman DG. Practical statistics for medical research. London: Chapman & Hall/CRC; 1991.

[14] Reed JF III. Homogeneity of Kappa statistics in multiple samples. Comput Methods Programs Biomed 2000;63:43–6.

[15] Stata/SE 8.0 for Windows. College Station, TX: Stata Corporation; 2003.

[16] Haley S, Harris SR, Tada WL, Swanson MW. Item reliability of the movement assessment of infants. Phys Occup Ther Pediatr 1986;61:21–39.

[17] Frankenburg WK, Dodds JB. The Denver Development Screening Test. J Pediatr 1967;71:181–91.

[18] Hammarlund K, Persson K, Sedin G, Stromberg B. A protocol for structured observation of motor performance in preterm and term infants. Interobserver agreement and intraobserver consistency. Ups J Med Sci 1993;98:77–82.

[19] Thomas SS, Buckon CE, Phillips DS, Aiona MD, Sussman MD. Interobserver reliability of the gross motor performance measure: preliminary results. Dev Med Child Neurol 2001;43: 97–102.

[20] Stuberg WA, White PJ, Miedaner JA, Dehne PR. Item reliability of the Milani-Comparetti Motor Development Screening Test. Phys Ther 1989;69:328–35.

[21] Gowland C, Boyce WF, Wright V, Russell DJ, Goldsmith CH, Rosenbaum PL. Reliability of the Gross Motor Performance Measure. Phys Ther 1995;75:597–602.

[22] Gross D, Conrad B. Issues related to reliability of videotaped observational data. West J Nurs Res 1991;13:798–803.

[23] Nordmark E, Hägglund G, Jarnlo GB. Reliability of the gross motor function measure in cerebral palsy. Scand J Rehab Med 1997;29:25–8.

[24] WHO Multicentre Growth Reference Study Group. Assessment of sex differences and heterogeneity in motor milestone attainment among populations in the WHO Multicentre Growth Reference Study. Acta Paediatr Suppl 2006;450:66–75.

[25] WHO Multicentre Growth Reference Study Group. WHO Motor Development Study: Windows of achievement for six gross motor development milestones. Acta Paediatr Suppl 2006;450:86–95.