University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

Faculty of Health: Medicine, Dentistry and Human Sciences School of Health Professions

2016-02-21

Rater reliability and scoring duration of the Quality Function Measure in ambulant children with hyperkinetic movement disorders

Marsden, JF

http://hdl.handle.net/10026.1/4991

10.1111/dmcn.13081
Developmental Medicine and Child Neurology
Wiley

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

This an accepted article published by Developmental Medicine and Child Neurology available at
http://onlinelibrary.wiley.com/doi/10.1111/dmcn.13081/abstract;jsessionid=F4E7EE0BF876CBA6ADC208C9E9E974DD.f04t03

DOI: 10.1111/dmcn.13081

Title: The Quality Function Measure: rater reliability and scoring duration when used in ambulant children with hyperkinetic movement disorders

Authors: Kylee Tustin[a], Hortensia Gimeno[a,b], Erin Morton[a], Jonathan Marsden[c]

Affiliations:

[a] Complex Motor Disorder Service, Evelina London Children's Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK

[b] King's College London, Institute of Psychiatry, Psychology Department, London, UK

[c] Plymouth University, Plymouth, UK

Correspondence to Kylee Tustin, Complex Motor Disorder Service, Evelina London Children's Hospital, Guy's and St Thomas' NHS Foundation Trust, Westminster Bridge Road, London SE1 7EH, UK.
E-mail: kylee.tustin@gstt.nhs.uk

Word Count Manuscript:	3050
Word Count Abstract:	200
Number of Tables:	2
Number of Figures:	2
Supplementary online figures:	3
References:	32

Aim: To examine intra- and inter-rater reliability/agreement, and time taken to score, when the Quality Function Measure (QFM) is applied to children with hyperkinetic movement disorders (HMD).

Method: Fifteen ambulant children with HMD participated (7 female; mean age 13y 7mo, SD 3y 7mo). Three trained raters (two physiotherapists, one occupational therapist) independently scored the QFM using videos of each child performing Gross Motor Function Measure (GMFM) Stand and Walk/Run/Jump dimensions. Reliability was evaluated using Intraclass Correlation Coefficient (ICC) model 2.1, Standard Error of Measurement (SEM) and Bland-Altman methods.

Results: Rater reliability was excellent for all five QFM attributes: intra-rater ICCs≥0.98 (95% CI 0.83-1.00), and inter-rater ICCs≥0.96 (95% CI 0.91-1.00). SEM varied from 2.07% to 4.72% points for intra- and inter-rater scores across QFM attributes. Bland-Altman tests demonstrated close agreement between ratings, with absolute mean differences varying from 0.34-3.23% (intra-rater) to 1.67-3.82% (inter-rater). Median scoring duration time was 83 minutes (range 56 to 144 minutes, SD 16.02).

Interpretation: Low measurement error attributable to rater effects suggests the QFM has potential as an evaluative measure in research studies involving children with HMD, though its lengthy scoring requirements are an important consideration for clinical practice. Evaluation of test-retest reliability and responsiveness is required.

What this paper adds:

- Criterion-trained raters are able to reliably use the QFM in children with HMD.
- The low measurement error related to rater effects suggests the QFM may potential as an evaluative measure in this patient group.
- Lengthy training and scoring demands suggest the QFM may be more viable in a research setting.

Hyperkinetic movement disorders (HMD) are associated with excessive involuntary movements including dystonia, chorea, athetosis, tremor and myoclonus[1]. Such disorders are seen frequently in children with neurological conditions, being associated with dyskinetic/dystonic cerebral palsy (CP) and numerous other congenital, acquired, and neurodegenerative conditions[2]. In addition to involuntary movements, paediatric movement disorders are often accompanied by multiple concomitant impairments, such as weakness and spasticity, which also contribute to disability[1].

Although diverse in aetiology and clinical presentation, the different movement disorders hold in common a disturbance of central motor control that manifests in alterations of posture and movement[2]. Along with restricted motor function, children may exhibit impairments in the quality of motor performance including disordered force and spatiotemporal characteristics[3], excess movement variability[4], postural instability, and malalignment[5]. Even in mild to moderate forms, children may have difficulty adapting their postural activity and/or motor behaviour to specific conditions or tasks[6]. While the impact of these motor (and other non-motor) impairments on activity and participation has not been fully explored, children and families report diverse concerns including pain, compromised volitional movement, and difficulties in daily functional activities and social participation[7].

Deep brain stimulation (DBS) is a neurosurgical option for paediatric movement disorders, though questions remain about its efficacy, particularly in secondary dystonias[8] where dystonia manifests as a symptom of an identified neurological insult or condition. Paediatric DBS has largely been evaluated using impairment/disease specific scales originally developed for adults, with the Burke-Fahn-Marsden Dystonia Rating Scale (BFMDRS) dominating in paediatric DBS literature. While such scales may incorporate judgments of functional ability, they primarily rate the presence and severity of involuntary movements and typically fail to discriminate among different movement disorders[9,10], or to consider the contribution of other concomitant impairments to functional difficulties[11]. Concerns have also been raised about the lack of sensitivity[9] and impact of age and development[10] when employing movement disorder rating scales in children. Such issues compromise the validity of judgments based on these instruments, particularly given DBS aims to ameliorate involuntary movements but has no direct effect on co-existing motor impairments.

For secondary dystonias in particular, discrepancies are noted in our own clinical practice[12,13], and in wider literature[8,11,14-16], between patient/family report and clinical assessment data, suggesting the scales in current use are not always adequately capturing changes following DBS in children. It is our impression that DBS may be associated with changes in the quality of motor performance, even in the absence of changes in standardised gross motor and dystonia rating scales. Research in children with CP suggests that improvements in quality of movement can be of practical and psychological significance to children and families[17,18]. This led us to seek clinical measures of movement quality that may be helpful in evaluating DBS outcomes in children.

The Quality Function Measure (QFM)[19] is a new observational criterion-referenced measure designed to evaluate gross motor movement quality in ambulant children

with CP. The scale, intended for use in tandem with the Gross Motor Function Measure (GMFM)[20], evaluates five quality attributes: Alignment, Coordination, Dissociated Movement, Stability, and Weight-shift, using the GMFM's Stand and Walk/Run/Jump dimension items. The distinction between these measures is that the GMFM evaluates "what" a child can do, while the QFM evaluates "how well" a child performs those same gross motor tasks.

Only a single validation study, involving children with predominantly spastic CP, is available for the QFM[19]. QFM scores differentiated children by Gross Motor Function Classification System (GMFCS) level[22], providing evidence of discriminant validity. Excellent rater and test-retest reliability were demonstrated for video-based rating, with coefficients ranging from 0.89 to 0.97. Minimal detectable change (MDC) estimates (9-11% points) suggest the scale has potential as an evaluative measure, with further work underway to establish responsiveness to change following intervention. Scoring required on average 66.7 minutes per child (range 15 to 180 minutes), with additional time (approximately 20 minutes) needed to compute test scores, highlighting the need to consider the scale's clinical utility.

Although these psychometric results are encouraging, reliability parameters are highly population specific and should be re-examined before a measure is applied to a novel patient population[23]. This is particularly pertinent given Wright et al (2014) included children with predominantly spastic CP[19]. Appreciable differences are seen in the clinical characteristics of hyperkinetic and spastic motor disorders, raising questions about the scale's applicability in HMD. Reliability testing was therefore considered important to ensure that repeated measures, undertaken by different clinicians, would agree sufficiently to allow between patient comparisons and identify true change in an individual where this occurs.

The current study therefore aimed to determine intra- and inter-rater reliability/agreement, and time taken to score, when the QFM is used in ambulant children with HMD in order to inform preliminary judgments about the scale's acceptability for clinical practice and research in this patient group.

METHOD

We conducted a reliability study comprised of intra- and inter-rater components, the latter including three replications per participant in a fully crossed design (i.e. all subjects were scored by the study's three raters). The National Research Ethics Service and the host hospital's Department of Research and Development approved the study. Written informed assent/consent was obtained from participating children and parents.

Participants

A convenience sample was drawn from consecutive children attending a specialist movement disorder service. Any young person was eligible providing they were (1) diagnosed with a HMD by a specialist paediatric neurologist; (2) aged 4-18 years; (3) able to walk, independently or with an assistive device; (4) considered to have sufficient language, understanding and attention to follow detailed study instructions. Children with significant lower limb spasticity (≥ grade 2 Modified Tardieu Scale[24]) were excluded.

A specialist paediatric neurologist classified each child's movement disorder on an aetiological basis[25] as primary dystonia (no neurological features other than dystonia, +/- tremor in some cases), dystonia-plus (inherited disorders whereby dystonia is accompanied by other neurological features including myoclonus and parkinsonism), secondary (movement disorder symptomatic of a neurological disorder or exogenous insult), or heredodegenerative (inherited disorders with evidence of brain degeneration). Functional mobility status was described using the GMFCS[22], with a "GMFCS equivalent" score applied to non-CP cases.

Raters

The study's raters (two physiotherapists [PTs] and one occupational therapist [OT]) all had at least two years paediatric movement disorder experience and prior experience with GMFM administration and scoring. QFM certification required raters to undertake a one-day in-person QFM training workshop and approximately 10 hours group scoring before independently passing the QFM video scoring criterion test administered by test developers.

Testing procedure

GMFM-66 dimensions D and E were administered by a trained physiotherapist (KT and EM), with items filmed using a standardised protocol intended to capture performance from both frontal and coronal planes of movement. Up to three trials were permitted for GMFM-66 scoring (as per manual guidelines), though only the first two trials were video-recorded for use in the QFM reliability study following information from test developers that reliability estimates were equivalent whether two or three item trials were performed (personal correspondence, V Wright 25.1.14). No shoes or orthotics were worn. Mobility aids were permitted for specific items as per QFM manual guidelines. Children continued through the test items according to individual ability.

Data collection

The QFM was scored for all participants using GMFM-66 videos. Individual GMFM-66 item trials were scored with reference to item-specific QFM criteria using a 4-point (0-3) ordinal scale, where '0' reflects "a lot of difficulty (markedly atypical)" and '3' represents "no difficulty (looks fine)". Each therapist independently rated all videos, documenting time required for scoring. No discussions were permitted among raters regarding participant scoring. Rater A scored all videos a second time after a minimum 2-week period, without reference to previous scores and ensuring at least two other study participants were rated during the intervening period to minimise potential recall bias. Only the first set of rater A scores (A1) contributed to inter-rater reliability analyses.

An Excel database (supplied by test developers) was used to calculate attribute summary scores from raw QFM item scores, and then convert these to percentage scores to adjust for the differing number of items per QFM attribute. A therapy assistant not involved in QFM rating processed raw test data from all raters.

Statistical analysis

Statistical analyses were performed using the Statistical Package for Social Sciences (version 21.0). Data were tested for normality using the Shapiro-Wilks test, with non-parametric approaches used where data were not normally distributed.

Descriptive statistics were calculated for QFM attribute summary scores and duration of scoring. Intra- and inter-rater reliability of QFM attribute scores was evaluated using ICC model 2:1, a two-way random effects single measures model of absolute agreement[26], with associated 95% confidence intervals (CI). General guidelines suggest that reliability coefficients of at least 0.9 are necessary to inform individual patient care, while coefficients of at least 0.7 are acceptable for group level research[27]. The ICC target was therefore 0.9, with 0.7 the lowest acceptable limit.

With three replications per subject, a sample of 12.8 participants was sufficient to test a hypothesised ICC of 0.9, with 0.7 as the lower acceptable limit (power = 0.80; $\alpha = 0.05$)[28].

Additional statistical methods were used to evaluate absolute agreement between different raters/ratings for the same subject. The standard error of measurement (SEM) was determined using the square root of the mean square error term from the ANOVA[29]. SEM values provide an indication of the absolute score differences needed before "true change" could be distinguishable from measurement error. Bland-Altman methods[30] were used to calculate the mean difference between ratings (μ_d), the standard deviation of the differences (SDdiff), and the 95% limits of agreement (LOA).

A paired samples t-test was applied to intra-rater scoring duration data to explore the potential impact of ongoing learning on time taken to score. As the number of video clips per child varied due to factors such as functional ability, filming issues and data capture, scoring duration was normalised for each case using the following equation:

$$Scoring\ duration\ \times \frac{Maximum\ 74\ trials}{Number\ of\ scoreable\ trials}$$

A Spearman's rank-order correlation was then used to assess the relationship between GMFCS level (or equivalent) and scoring duration, using both raw and normalised scoring data.

RESULTS

Fifteen children (7 female), with a mean age of 13 years 7 months (SD 3 years 7 months) participated in the study. Aetiology and functional mobility status varied widely (Table SI, supporting information published online). Five participants had primary dystonia (largely genetically undetermined), six dystonia-plus movement disorders, and four dyskinetic CP. Six participants did not require mobility aids, five used wheelchairs to support distance mobility, while four utilised walking aids for indoor mobility.

Mean QFM attribute scores, based on pooled results from all ratings, varied from 50.7% (SD 23.3) for alignment to 64.1% (SD 23.4) for dissociated movement. GMFM-66 total score estimates varied from 60.9 to 100, with a mean of 80.9 (SD 13.6) (Table SII, supporting information published online).

Rater reliability

As shown in table 1, both intra- and inter-rater ICC estimates were excellent for all QFM attributes (range 0.96-0.99), the lower CI limit exceeding 0.90 for all except intra-rater coordination scores (lower CI 0.83). SEM values were 2.1-3.7% points for intra-rater scores and 2.1-4.7% points for inter-rater scores.

Bland-Altman tests demonstrated close agreement between all ratings, with absolute mean differences varying from 0.3-3.2% points (intra-rater) to 1.7-3.8% points (inter-rater). Bland-Altman plots (not provided) suggested no obvious relationship between measurement error and the measured value. Despite the small sample size, the LOA for intra-rater test scores did not exceed +/-10%, with the exception of Alignment (upper 95% CI limit 11.7%). Inter-rater agreement was more variable. Absolute differences between ratings for the same item in the same participant ranged from perfect agreement to 21.4% points (Dissociated Movement), though absolute differences exceeded 10% in only 8% (18/225) of rater pairings. Statistically significant between-rater differences were found for Dissociated Movement ($F(14,28)=4.2$, $p=0.026$) and Stability ($F(14,28)=7.6$, $p=0.002$). A plot of subject results by rater for these attributes (Figures 1 and 2) demonstrates highly congruent ratings, though suggests a tendency for rater B to allocate slightly lower scores than other raters. This tendency appears more evident in primary/dystonia-plus disorders (participants 1-8).

Figure 1

Figure 2

Bland-Altman tests (Table II) for these QFM attributes confirm this pattern, with rater B allocating, on average, slightly lower scores (1.9-3.8% points) than both rater A and C, while the mean difference between raters A and C was close to zero (0-1.1% points).

Scoring duration

The median time required for video-based scoring was 83 minutes (SD 16.02), varying from 56 minutes for a GMFCS level I equivalent participant to 144 minutes for a level II child, both of whom completed 35 test items. Differences in scoring duration among raters were not significantly different, Friedman test $\chi2(2)=3.6$, $p=0.165$.

A significant reduction in scoring duration was seen across the two intra-rater scoring rounds, with a mean difference of 13.9 minutes (95% CI 6.5 - 21.2, $p=0.001$) and absolute differences across the two rounds ranging from -1 to -49 minutes.

Although children classified as GMFCS level III equivalent were able to complete fewer tasks, they required a similar time for scoring as those in other GMFCS levels who completed a greater number of test items. No significant relationship was seen between GMFCS level and time taken to score using raw scoring duration data. However, after correcting for the variable number of trials between children, a positive correlation was seen between GMFCS level and normalised scoring duration for all raters (Table SIII, supporting information published online).

DISCUSSION

Evidence-based best practice requires healthcare providers to quantitatively demonstrate the efficacy and effectiveness of their interventions. This demands measures that are accurate, reliable and responsive to clinically important change.

When clinical measures require judgments on the part of human raters, rater reliability becomes essential to the validity of data[31].

This study was undertaken to provide preliminary psychometric evidence for the use of the QFM in children with HMD. Intra-and inter-rater reliability for QFM attribute summary scores was found to be excellent (ICC point estimates ≥0.96), with small SEM values and Bland-Altman test results providing further evidence that measurement error related to rater effects is acceptably low. This has implications for both clinical practice and research, as robust rater reliability is necessary for interpreting individual patient change, while strong reliability coefficients enhance statistical power and minimise sample size requirements for future studies.

Although these results suggest the QFM shows promise as an evaluative measure of gross motor performance quality in this patient group, it's clinical utility is limited by the time and training required to achieve competence, the lengthy scoring burden, and the scale's limited applicability to ambulant patients. The tool may be more viable in a research context, where support for filming, video editing, scoring, and data processing can be secured in advance.

From an interpretation viewpoint, it is not clear whether the longer scoring duration for children with greater disability related to their requiring more time to execute the tasks, such that a longer video sequence needed to be viewed, or whether raters found it more difficult to assign scores to children with greater functional impairment. However, the evidence of reduced scoring duration at retest, as evaluated with rater A, suggests familiarisation with scoring guidelines may offer relative improvement in scoring duration over time. In view of the substantial scoring burden, intra-rater reliability results are based on a single rater. Having all raters score videos a second time would have enhanced the validity of these results. Although it is not clear whether the differences seen between raters in the present study would be clinically important, consistently closer agreement was seen between raters A and C, suggesting that certain raters may be more interchangeable than others. Rater B tended to allocate slightly lower scores, at least for the attributes of Dissociated Movement and Stability. This tendency appeared more evident in primary/dystonia-plus disorders (participants 1-8), raising the possibility that some QFM attributes may be more challenging to evaluate in these movement disorders. A larger sample size would be required to allow valid comparison of rater reliability across different aetiological sub-groups.

The raters came from different disciplines (PT and OT) and ranged in post-qualification experience (range 4-16 years). All had at least 2 years' experience working within a specialist paediatric movement disorder service, including administering and scoring the GMFM in patients with HMD. These are all factors that may influence gross motor observational skills. Although it cannot be assumed that comparable reliability estimates will be achieved using raters with a different clinical background, our results are in line with those of Wright et al. (2014), who utilised physiotherapist raters working across several centres[19]. Nonetheless, it is recommended that reliability data be replicated with a broader range of raters recruited from multiple movement disorder centres, in conjunction with test-retest reliability estimates and subsequent evaluation of test responsiveness, to allow a

more definitive conclusion to be made about the QFM as an evaluative measure for children with HMD.

The current study evaluates reliability of video-based scoring using participant data from a single point in time. It does not evaluate the impact on reliability of differences in test administration (i.e. if the same or different therapists were to apply the test on different occasions) or a child's session-to-session variability. Test-retest data would be needed to determine the difference in scores required to represent "true change" in a child's motor performance. Further, before test responsiveness is formally evaluated, it is important to ascertain the extent to which motor performance varies day-to-day in this patient group despite a stable management strategy. If performance fluctuates markedly, this added "noise" is likely to present considerable difficulties in demonstrating change following DBS in secondary dystonia, where the intervention effect is anticipated to be moderate to small[15].

Provided psychometric properties are shown to be acceptable in future studies involving children with HMD, the GMFM and QFM together may allow us to systematically evaluate the nature of gross motor response to DBS in ambulant children. It is possible that DBS offers a smoothing or stabilisation of performance in some children rather than the acquisition of new gross motor skills. Establishing consistency of performance, ideally at the upper limit of capability, may well be perceived as meaningful, even in the absence of measureable change in gross motor function. It is also possible that improvement in qualitative aspects of gross motor function, such as stability, coordination and effort, may positively influence manual function and ability to execute activities of daily living (ADL), though such associations require evaluation. If inherent performance variability was indeed demonstrated in this patient group, multiple baseline research methodologies could be used to explore the extent to which DBS stabilises motor variability and alters underlying motor capacity. Further, it is important to determine the degree to which treatment responses differ depending on aetiology or movement disorder phenotype, and to evaluate relationships between changes in movement quality, gross motor function and patient-rated goal attainment. The QFM, as part of a broad assessment across all three domains of the International Classification of Function, Disability and Health framework[32], may allow us to elucidate the extent to which DBS has a clinically meaningful effect on gross motor and ADL performance, particularly in children with secondary movement disorders where patient selection and treatment efficacy remain topics of considerable debate.

ACKNOWLEDGEMENTS

Thanks must first go to the families and young people who participated in this project. Acknowledgement is given to Holland Bloorview Kids Rehabilitation Hospital, Canada, for providing QFM materials and training. Special thanks are given to Doctor Virginia Wright, who provided QFM training and the logistical support required to share the QFM for this project, and who gave helpful comments on earlier drafts of this paper. Thanks are also owed to Oliver Ewing for videotaping, data processing and administrative support. Kylee Tustin gratefully acknowledges

the financial support for post-graduate tuition fees provided by Guys & St Thomas' NHS Foundation Trust. The Trust was otherwise not involved in any aspect of the study design, data collection, data analysis or manuscript preparation.

REFERENCES

1. Sanger TD, Chen D, Fehlings DL, Hallett M, Lang AE, Mink JW, et al. Definition and classification of hyperkinetic movements in childhood. Mov Disord. 2010;25(11):1538-49.
2. Delgado MR, Albright AL. Movement Disorders in Children: Definitions, Classifications, and Grading Systems. J Child Neurol. 2003;18(1 suppl):S1-S8.
3. Malfait N, Sanger TD. Does dystonia always include co-contraction? A study of unconstrained reaching in children with primary and secondary dystonia. Exp Brain Res. 2007;176(2):206-16.
4. Pavone L, Burton J, Gaebler-Spira D. Dystonia in Childhood: Clinical and Objective Measures and Functional Implications. J Child Neurol. 2013;28(3):340-50.
5. Boyce W, Gowland C, Russell D, Goldsmith C. Consensus methodology in the development and content validation of a gross motor performance measure. Physiother Can. 1993;45:94-.
6. Hadders-Algra M. The neuronal group selection theory: promising principles for understanding and treating developmental motor disorders. Dev Med Child Neurol. 2000;42(10):707-15.
7. Gimeno H, Gordon A, Tustin K, Lin J-P. Functional priorities in daily life for children and young people with dystonic movement disorders and their families. Eur J Paediatr Neurol. 2013;17 (2):8.
8.
9. Monbaliu E, Ortibus E, Roelens F, Desloovere K, Deklerck J, Prinzie P, et al. Rating scales for dystonia in cerebral palsy: reliability and validity. Dev Med Child Neurol. 2010;52(6):570-5.
10. Mink JW. Special concerns in defining, studying, and treating dystonia in children. Mov Disord. 2013;28(7):921-5.
11. Koy A, Pauls KAM, Flossdorf P, Becker J, Schönau E, Maarouf M, et al. Young Adults with Dyskinetic Cerebral Palsy Improve Subjectively on Pallidal Stimulation, but not in Formal Dystonia, Gait, Speech and Swallowing Testing. Eur Neurol. 2014;72(5-6):340-8.
12. Gimeno H, Tustin K, Selway R, Lin J-P. Beyond the Burke–Fahn–Marsden Dystonia Rating Scale: Deep brain stimulation in childhood secondary dystonia. Eur J Paediatr Neurol. 2012;16(5):501-8.
13. Gimeno H, Tustin K, Lumsden D, Ashkan K, Selway R, Lin J-P. Evaluation of functional goal outcomes using the Canadian Occupational Performance Measure (COPM) following Deep Brain Stimulation (DBS) in childhood dystonia. Eur J Paediatr Neurol. 2014;18(3):308-16.
14.
15. Koy A, Hellmich M, Pauls KAM, Marks W, Lin J-P, Fricke O, et al. Effects of deep brain stimulation in dyskinetic cerebral palsy: A meta-analysis. Mov Disord. 2013;28(5):647-54.
16.
17. Buckon CE, Thomas SS, Piatt Jr JH, Aiona MD, Sussman MD. Selective dorsal rhizotomy versus orthopedic surgery: a multidimensional assessment of outcome efficacy. Arch Phys Med Rehabil. 2004;85(3):457-65.
18. Eliasson A-C, Öhrvall A-M, Borell L. Parents' Perspectives of Changes in Movement Affecting Daily Life Following Selective Dorsal Rhizotomy in Children with Cerebral Palsy. Phys Occup Ther Pediatr. 2000;19(3-4):91-109.

19. Wright F, Rosenbaum P, Fehlings D, Mesterman R, Breuer U, Kim M. The Quality Function Measure: reliability and discriminant validity of a new measure of quality of gross motor movement in ambulatory children with cerebral palsy. Dev Med Child Neurol. 2014;56(8):770-778
20. Russell DJ, Rosenbaum PL, Wright M, Avery LM. Gross Motor Function Measure (GMFM-66 and GMFM-88) User's Manual. 2nd Edition ed. Ontario, Canada: Mac Keith Press; 2013.
21.
22. Palisano RJ, Rosenbaum P, Bartlett D, Livingston MH. Content validity of the expanded and revised Gross Motor Function Classification System. Dev Med Child Neurol. 2008;50(10):744-50.
23. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud. 2011;48(6):661-71.
24. Boyd RN, Graham HK. Objective measurement of clinical findings in the use of botulinum toxin type A for the management of children with cerebral palsy. Eur J Neurol. 1999;6:s23-s35.
25. Geyer HL, Bressman SB. The diagnosis of dystonia. Lancet Neurol. 2006; 5(9):780-90.
26. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.
27. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use [e-book]. Oxford : : Oxford University Press; 2008. Available from: http://kcl.eblib.com/patron/FullRecord.aspx?p=665468.
28. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17(1):101-10.
29. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical Methodology for the Concurrent Assessment of Interrater and Intrarater Reliability: Using Goniometric Measurements as an Example. Phys Ther. 1994;74(8):777-88.
30. Bland MJ, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327(8476):307-10.
31. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 3rd ed. New Jersey: Prentice-Hall Health; 2009.
32. World Health Organisation (WHO). ICIDH-2: International Classification of Functioning, disability, and health. 2001 [4.2.12]; Available from: www.who.int/classification/icf

Table I: Quality Function Measure rater reliability statistics

Intra rater reliability - QFM attribute summary scores

Attribute	ICC (2.1)	95% CI	SEM
Alignment	0.98	0.94 - 0.99	3.7
Coordination	0.99	0.83 - 1.00	2.1
Dissociated Movement	0.99	0.97 - 1.00	2.6
Stability	0.99	0.97 - 1.00	2.6
Weight shift	0.99	0.96 - 1.00	2.9

Inter rater reliability - QFM attribute summary scores

Attribute	ICC (2.1)	95% CI	SEM
Alignment	0.98	0.96 - 0.99	3.0
Coordination	0.98	0.94 - 0.99	3.7
Dissociated Movement	0.96	0.91 - 0.99	4.2
Stability	0.99	0.97 - 1.00	2.1
Weight shift	0.96	0.92 - 0.99	4.7

Legend: CI = confidence interval; ICC = intraclass correlation coefficient (single measures, absolute agreement); SEM = standard error of measurement.

Table II: Bland-Altman test results for inter-rater Quality Function Measure attribute summary scores

QFM attribute	μ_d	SDdiff	Lower LOA	Upper LOA	Min abs diff	Max abs diff
Alignment						
Rater A1 and B	1.4	4.9	-8.3	11.2	0.0	8.3
Rater C and B	2.2	4.0	-5.8	10.2	0.0	8.3
Rater A1 and C	-0.8	3.9	-8.5	6.9	0.0	9.7
Coordination						
Rater A1 and B	3.0	6.7	-10.5	16.4	0.5	16.7
Rater C and B	2.8	5.8	-8.8	14.4	0.0	13.2
Rater A1 and C	0.2	2.0	-3.7	4.1	0.0	3.4
Dissociated Movement						
Rater A1 and B	3.8	6.1	-8.4	15.9	1.2	15.5
Rater C and B	3.8	7.6	-11.4	19.1	0.0	21.4
Rater A1 and C	0.0	3.0	-6.0	6.0	0.0	6.0
Stability						
Rater A1 and B	3.0	3.4	-3.8	9.8	0.0	10.2
Rater C and B	1.9	3.5	-5.2	9.0	0.0	9.1
Rater A1 and C	1.1	1.9	-2.7	4.8	0.0	6.5
Weight shift						
Rater A1 and B	-1.7	7.7	-17.0	13.6	0.7	17.5
Rater C and B	-0.9	7.3	-15.5	13.8	0.7	13.5
Rater A1 and C	-0.8	4.6	-10.0	8.4	0.0	11.1

Legend: abs diff = absolute difference; LOA = limits of agreement; SDdiff = standard deviation of the differences; μd = mean difference between allocated scores

Figure 1: Quality Function Measure Dissociated Movement subject by rater scores

Figure 2: Quality Function Measure Stability subject by rater scores