Taylor & Francis
Taylor & Francis Group

Disability and Rehabilitation

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/idre20

Reliability and minimal detectable change of the *Challenge*, an advanced motor skills test for children with cerebral palsy, Danish version

Kirsten Nordbye-Nielsen, Thomas Maribo, F. Virginia Wright, Ole Rahbek & Bjarne Møller-Madsen

To cite this article: Kirsten Nordbye-Nielsen, Thomas Maribo, F. Virginia Wright, Ole Rahbek & Bjarne Møller-Madsen (2022) Reliability and minimal detectable change of the *Challenge*, an advanced motor skills test for children with cerebral palsy, Danish version, Disability and Rehabilitation, 44:16, 4485-4492, DOI: 10.1080/09638288.2021.1906332

To link to this article: https://doi.org/10.1080/09638288.2021.1906332

Published online: 06 May 2021.

DISABILITY AND REHABILITATION
2022, VOL. 44, NO. 16, 4485–4492
https://doi.org/10.1080/09638288.2021.1906332

Taylor & Francis
Taylor & Francis Group

ASSESSMENT PROCEDURE

Reliability and minimal detectable change of the *Challenge*, an advanced motor skills test for children with cerebral palsy, Danish version

Kirsten Nordbye-Nielsen[a,b], Thomas Maribo[a,c], F. Virginia Wright[d,e], Ole Rahbek[f,g] and Bjarne Møller-Madsen[a,b]

[a]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; [b]Department of Children's Orthopaedics, Aarhus University Hospital, Aarhus, Denmark; [c]DEFACTUM, Central Region Denmark, Aarhus, Denmark; [d]Department of Physical Therapy, University of Toronto, Toronto, Canada; [e]Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada; [f]Department of Clinical Medicine, Aalborg University, Aalborg, Denmark; [g]Department of Childrens's Orthopaedics, Aalborg University Hospital, Aalborg, Denmark

ABSTRACT

Purpose: To translate and cross-culturally adapt the *Challenge*, and investigate the reliability and minimal detectable change (MDC) of the Danish *Challenge* in children with cerebral palsy (CP).

Materials and methods: A Danish version of the *Challenge* was created through a standardized translation process. Four physiotherapists evaluated face validity. Independently ambulatory children with CP were tested. Live performance rating was conducted by assessors independently scoring the *Challenge*. Video-rating was undertaken for a subset of assessments. Same day assessment test–retest reliability was estimated. The *Challenge*'s Best Score Total was of primary interest.

Results: Forty-five children (5–18 years: mean 10 years 9 months; 19 girls) in Gross Motor Function Classification System levels I and II were tested. Inter-rater reliability was excellent for live assessments ($n = 45$) ICC = 0.998 (95% CI 0.998–0.999) and video assessments ($n = 15$) ICC = 0.991 (95% CI 0.963–0.997) and intra-rater reliability was excellent for live versus video-recorded assessments ($n = 10$) ICC = 0.977 (95% CI 0.895–0.994). Test–retest reliability ($n = 22$) was excellent with ICC = 0.991 (95% CI 0.979–0.996) and minimal detectable change (MDC_{90}) of 4.7 points.

Conclusions: The Danish *Challenge* showed excellent reliability in this testing context when physiotherapists scored from live- or video-recorded assessments. The *Challenge*'s ability to detect 4.7 points change seems a clinically realistic target for progress.

Clinical trial registration: This trial has been approved by the Data Protection Agency, Central Region Denmark, Ref nr.: 615216, Case nr.: 1-16-02-46-16. Registration date: 01-01-2016.

➤ IMPLICATIONS FOR REHABILITATION

- The *Challenge* remained reliable and maintained a promising minimal detectable change of less than five points after translation and cultural adaptation.
- The Danish version of the *Challenge* 20-item version can be used to measure advanced motor skill performance in children with cerebral palsy, GMFCS level I and GMFCS level II.
- *Challenge* live scoring is as reliable as the more time-consuming video-recorded scoring, meaning that physiotherapists can choose the method that fits best with their clinical context and preference.

ARTICLE HISTORY
Received 30 September 2020
Revised 14 March 2021
Accepted 17 March 2021

KEYWORDS
Gross motor function; psychometric properties; disabilities; ambulatory children; cerebral palsy

Introduction

Cerebral palsy (CP) is as an umbrella-term for a group of disorders causing motor disability in children [1]. Physical function is affected and typically influences abilities in physical activities and participation with peers [2,3]. Children with CP often receive medical and rehabilitation interventions to maximize gross motor function and prevent secondary musculoskeletal and functional deterioration [4]. Planning and evaluation of these interventions, using well-targeted and psychometrically sound outcome measures, is essential [4,5].

The Gross Motor Function Classification System (GFMCS) is widely used to classify children, and helps clinicians optimize the selection of best-fit interventions. As estimated in high-income countries including Denmark, about two-thirds of children with CP are ambulant with the highest proportion in GMFCS level I and level II [6–8], meaning that they are ambulatory and independent in walking and daily activities, but have limitations of coordination, balance, and speed [9]. They often receive physiotherapy to improve physical function focusing on enhancing advanced gross motor skills for participation in sports and recreation-based activities [8,10,11]. When thinking of the impact of gross motor function interventions, it is valuable to include a measure that permits evaluation of performance abilities as reflected by coordination, balance, and speed in activities involving upper and lower extremities [12,13].

The Gross Motor Function Measure (GMFM) is an outcome measure designed to evaluate gross motor function in children with CP at all GMFCS levels [14,15]. However, it was not

CONTACT Kirsten Nordbye-Nielsen kirstennordbye@clin.au.dk Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, 8200 Aarhus N, Denmark

Table 1. Description of categories and items and tasks of the *Challenge*[a].

Category	Item no.	Task	Included in the 20-item
Balance/coordination	2	Catch and throw a ball four cycles	Yes
	3	Bounce a basketball (10 times)	No
	4	Throw tennis ball in a target	No
	5	Bounce a tennis ball (5 times, both hands)	Yes
	6	Run and kick a soccer ball down path	No
	7	Walks sideways and return on 5-meter line	Yes
	8	Step sideways over stick (4 times)	Yes
	19	Single leg stance (20 s both legs)	Yes
	20	Tandem stance (20 s)	No
	24	Step in and out of lines (5 times)	Yes
	25	Walk in a wooden beam, controlled stop	Yes
Walk/run/jump	1	Star Jumps (10 times)	Yes
	9	Walk, turn, and walk backwards in path	Yes
	10	Run in path and controlled stop on end line	Yes
	11	Run, pick up pin, and run back in path	Yes
	12	Run weaving through pylons (6 pylons)	Yes
	13	Walk backwards on line (3 m)	Yes
	14	Jump forward, controlled landing	Yes
	15	Skip forward down path (no rope)	No
	16	Jumps with a skipping rope	Yes
	22	Step up and down (5 cycles both legs)	Yes
	23	Sideways jumps across line (5 meters)	Yes
Dual task	17	Walk with a lunch tray and glass down path	Yes
	18	Walk bouncing a basketball	Yes
	21	Dribble a soccer ball down path	Yes

[a]*Challenge*: the 25 items version presented with category, item numbers and task descriptions, and items included in the 20-item version.

developed to measure advanced motor skills, and shows a ceiling effect when used to measure foundational skills in children in GMFCS I aged five years and older [15–18]. Use of well-known norm-based advanced motor measures such as the Movement ABC and Bruininks-Oseretsky Test of Motor Proficiency is not appropriate to evaluate change over time in children with disabilities as they tend to fall further behind on the development curve as they age [19]. Hence, clinicians have lacked the ability to measure changes in advanced gross motor performance in these children, meaning that we cannot know what the physical impact is of interventions that are targeted toward improvement of skills that underlie participation in sports and recreation-based activities. A comprehensive evaluation of gross motor performance requires standardized assessment tools, with acceptable validity and reliability in the target population [20] in this case, children with CP.

The *Challenge* is a new observational measure developed to fill this measurement gap [21, 22]. Its basic psychometric properties have been established with excellent inter-rater and test–retest reliability [22]. Discriminant validity has been demonstrated with respect to children in GMFCS I versus level II [23]. As far as concurrent validity, there was strong association between the *Challenge* and the Test of Gross Motor Development-2 ($r = 0.76$) and also with four single skill tests, i.e., 10×5 m Sprint Test ($r = -0.82$), the Muscle Power Sprint Test ($r = -0.71$), and vertical/broad jump distances ($r = 0.77$) [23]. Rasch scaling has been accomplished, showing the *Challenge* to be a unidimensional scale, and its items are harder than the most difficult of the GMFM-66 skills [24]. Initial use within published intervention studies has shown mean gains of 2.8 points and median gains of 4.5 points in association with a sports skills intervention program [25] and a therapist-monitored home active video gaming (AVG) program, respectively [26]. Drawing from this evidence, in Clutterbuck et al.'s systematic review of sports-focused high level gross motor assessments for ambulatory children with CP [23], the *Challenge* was recommended in the measurement selection decision tree as a CP-specific tool with promising psychometric properties and good clinical utility in the area of technique, speed, and accuracy.

Prior to implementation of a new measure for children in a different language and context, its psychometric properties need to be examined. Translation of the *Challenge* into Brazilian-Portuguese, followed by evaluation in a sample of children ages 5–18 in GMFCS I and II demonstrated excellent reliability, validity, and acceptable responsiveness to change over time in that context [27]. Translation into other languages will help to expand the *Challenge*'s valid use across a wider international group of clinicians and children.

Therefore, this study's purpose was to investigate the reliability of the Danish-translated version of the *Challenge* in ambulatory children with CP age 5–18 years. The objectives were to: (1) estimate inter- and intra-rater reliability among trained physiotherapist assessors for live assessments, and video scoring contexts, (2) same day test–retest reliability, and (3) examine the *Challenge*'s minimal detectable change (MDC). The hypothesis was that the *Challenge* is a reliable tool to assess advanced gross motor function with an MDC_{80} of less than five points.

Methods

Translation of the challenge

The *Challenge* (20-item version) [24] was used with permission from its developer FV Wright. Translation into Danish was performed according to the guidelines from WHO as described by Beaton et al. [28,29] as follows: (1) translation by two independent translators (T1, T2). T1 was a physiotherapist specialized in pediatrics and the principal investigator, and T2 was a linguistic professional translator without specific knowledge on the construct and subject area; (2) synthesizing the translations, in order to achieve coherence; (3) face validity evaluation on clear wording and importance (yes/no) by four physiotherapists in CP; and (4) English back-translation of the consensus version by a

professional translator (T3) without disease specific knowledge. The developer reviewed and responded with linguistic comments and final revisions were then made to the Danish *Challenge*.

Setting and design

Reliability testing of the *Challenge* was performed at Aarhus University Hospital, children from the outpatient clinic were invited and informed consent obtained from parents. Assessments took place at individual appointments during afternoons, weekend, or holidays to facilitate participation of families. The study was approved by the Danish Data Protection Agency and notified to the local ethics committee.

Participants

Children were included if they: (1) had confirmed diagnosis of CP; (2) were in GMFCS level I or II; and (3) were five to 18 years inclusive. Children were excluded if they: (1) used a gait aid or (2) a parent, after reading the study information letter, was of the impression that their child could not follow the instructions required to perform the *Challenge* in one same day session.

Measure

The *Challenge* measures performance ability related to coordination, accuracy and speed of 20 items of advanced motor skills [22,24]. It consists of three motor skill categories: (1) balance/coordination, (2) walk/run/jump, and (3) dual task. Five items were removed from the 25 items *Challenge* version during the Rasch scaling process [24], specifically item 3: bounce a basketball (10 times), item 4: throw tennis ball in a target, item 6: run and kick a soccer ball down path, item 15: skip forward down path (no rope), and item 20: tandem stance (20 s) (Table 1) to create the *Challenge-20* that was then used in this study. The *Challenge* aims to elicit the child's best performance abilities within a supportive test situation with each item tested three times using a dynamic assessment style. Item performance directions (i.e., difficulty) are systematically adapted as needed to suit the abilities of the child and give them opportunity to demonstrate their best performance. Engagement guidelines developed for the *Challenge* testing procedure illustrate how this is done [30].

Scoring is on a five-point scale for which certain item-specific behavioral achievements are required. This scale measures the child's ability to perform the skill (scores of "0" to "2") and their performance accuracy and speed (scores of "3" and "4"). Children aged 12 years and up with no motor disabilities are typically developing are able to score "4" on most or all items in the *Challenge* (Personal communication; FV Wright). A cumulative total score (percentage) is calculated from each of the child's best trial item scores (primary score), first trial item scores, and mean trial item scores [22].

Reliability study procedure

Inter-rater, intra-rater and test–retest reliability evaluation of the Danish *Challenge* was carried out for live and video-recorded assessments. Two of four trained study assessors independently scored each assessment. Live assessments of children took place on and around the *Challenge's* pathway (0.45 m × 10 m) located in a quiet hallway. Setup for each item, engagement guidelines, use of standardized testing materials and scoring were as outlined in the *Challenge* manual.

Study assessors

Three assessors A, B, and C were involved in administering the *Challenge* assessments and scoring the live assessments. Assessor D scored only the video-recorded assessments. Assessors A and D were physiotherapists experienced in working with children with CP, while assessors B and C were physiotherapists with no previous experience in this area. The intention with this breadth in experience was to have the assessors in some way reflect the diversity of experience of those who expected to use the *Challenge* in clinical practice. All were trained on the *Challenge* and passed a criterion test prior to beginning study assessments. Criterion training requirements included understanding the *Challenge* training materials and engagement guidelines, and successfully scoring two test videos to 90% accuracy.

Assessor A administered the live assessments with all the children. Using the *Challenge* test process, each child watched the assessor demonstrate the item, had a practice trial, and was given three test trials unless s/he scored the maximum four points (at which point no further trial of the item was done) or chose not to repeat the item. If a child decided not to try an item, this was respected as part of the supportive testing style, although they were given a score "0" for that item based on the conservative scoring assumption that refusal meant they felt they could not do it.

Assessor A, and either assessor B or C as available, independently scored the 45 children in the inter-rater reliability evaluation of live assessment. In separate inter-rater evaluation, assessor B (no experience in CP) and assessor D (with experience in CP), scored the videos from children who had video-recorded tests. Intra-rater reliability from assessor B_{live} versus B_{video} was evaluated. Finally, for evaluation of test–retest reliability (live performance), assessor A re-scored a subsample of 22 children who were tested twice on the same day with a 2–3 h break between tests.

Sample size

Based on sample size recommendations for valid interpretation of ICC's [31], a sample of 45 children was planned for the reliability analyses with the enrolment goal of at least 30 children for the test–retest portion of the study. These participant numbers are in line with that of the original *Challenge* reliability study [22] as well as other reliability studies of pediatric fine and gross motor function measures, i.e., from 25 to 50 participants [15,22,32–35].

Statistical analysis

Descriptive statistics were calculated for the *Challenge* best score, as well as the first and mean scores. To evaluate inter-rater, intra-rater and test–retest reliability, intraclass correlations ICC's (type 2,1) [36] two-way-random analysis [37], and associated confidence interval's (95% CI) and standard error of measurement (SEM) were estimated. Rater agreement within three test scenarios was also examined by a Bland–Altman plot [38]. MDC was calculated at the 80% and 90% (MDC_{80} and MDC_{90}) levels for test–retest data to give an estimate of score difference reflecting change beyond error. Sub-analysis within GMFCS levels was conducted. Statistical analyses were performed using Stata 16 (StataCorp, College Station, TX).

Results

Translation

The face validity evaluation of the translated version revealed the need for a few changes in wording without changing the meaning. For example, in item 2, "or a bouncy ball" in translation into Danish would be " a small ball not equivalent to a basketball", and translated to "or a ball equivalent to a basketball" so as to ensure the correct size. In item 21, the English task description is that a child should use "foot to foot pass style", which to ensure its meaning in Danish was translated into a single Danish equivalent word. The four physiotherapists answered "*yes*" on relevance of all items and agreed on the revised wording, and the back-translated version was accepted by the developer of the *Challenge*.

Reliability

Forty-five children with CP, GMFCS level I or II (age 5–18 years, mean 10.9(4.0)) were included (Table 2). Two raters assessed all children during the live assessment, 15 children were video-recorded for the video versus video and the live versus video ratings, and 22 children performed two tests on the same day for test–retest reliability (Table 3).

Reliability estimates were excellent (ICC >0.90) for the different rating scenarios for best, first, and mean score totals with 95% CI >0.78, with lowest estimate of the lower CI being for the first score total in the live versus video rating scenario (Table 4). The SEM estimates (Table 4) varied from 0.87 for inter-rater (live scoring) best score total to 5.20 for the live test–retest first score total results. In the inter-rater sub-analysis, the consistency of assessors B and C with assessor A for their portion of the reliability sample was evident (Table 4).

Table 2. Participant characteristics; number and age.

	Number of children $n=45$	GMFCS[a] Level I $n=25$	GMFCS Level II $n=20$
Children			
Girls	19	10	9
Boys	26	15	11
Age, years			
5–8	16	12	4
9–12	13	3	10
13–18	16	10	6
Mean age (SD)	10.9 (4.0)	10.5	11.4
Age range	5–18	5–18	8–17

[a]GMFCS: Gross Motor Function Classification System.

The Bland–Altman plots revealed no evidence of measurement bias for inter-rater live assessment scenario ($n=45$) (Figure 1(a)). With the comparison of inter-rater reliability (video) for assessor B versus D ($n=15$), there was a shift midway in the direction of difference (albeit small) between the two with assessor B scoring lower than assessor D for scores below 40% and then higher than D for scores from 40% up. For the intra-rater assessor B (video vs. live) ($n=10$) and the test–retest assessor A (live) ($n=22$), there was some indication of greatest differences in first and second ratings at the *Challenge*'s 40–50% point scoring range.

MDC estimates were as follows: best score total MDC_{80}=3.7, and MDC_{90}=4.7, first score total MDC_{80}=9.4 and MDC_{90}=12.1, and mean score total MDC_{80}=5.6 and MDC_{90}=7.2. The targeted MDC_{80} of <5 points was achieved for the best score total (primary score), but not for first or mean score totals.

Discussion

This study investigated the Danish *Challenge*'s ability to consistently measure advanced motor skills with independently ambulatory children with CP. Translation from English to Danish and back was carried out according to international guidelines for cross-cultural adaptation of health-related measures by Beaton et al. [28,29]. In this process, only minor adaptations to the original version were required to convert the English version into Danish, to keep the semantic meaning while reflecting the cultural context.

The Danish *Challenge* retained excellent inter-rater, intra-rater, and test–retest reliability, and results were consistent with reliability estimates in the original *Challenge* study [22]. We assessed the reliability from video-recorded assessment in a sub-sample to see if scoring from video-recordings was superior to live scoring in light of the chance to view a child's item performance more than once during video review. The results revealed similar reliability estimates (ICC's > 0.90) for each approach. Reliability of video-recorded *Challenge* assessments has not previously been evaluated in children with CP. However, a study using the companion Acquired Brain Injury Challenge Assessment that evaluated video-recorded assessments with children with brain injury also showed excellent reliability estimates (ICC's > 0.90) that were comparable to live rating [39]. As well, the comparability of reliability scoring for video-recorded versus live assessment was also seen in studies of the GMFM assessments in Brazilian children with CP [40]. From a clinical point scoring, live assessment is likely preferable as it is more time and cost efficient, i.e., beneficial to do it all in "one round".

Table 3. *Challenge* total scores.

	n	Best score Mean (SD)	Range	First score Mean (SD)	Range	Mean score Mean (SD)	Range
Inter-rater: score							
Assessor A	45	48.89 (24.30)	11.96–92.39	39.25 (20.79)	9.78–78.26	42.36 (21.84)	11.23–80.98
Assessor B + C	45	48.93 (24.04)	13.04–92.39	39.57 (20.55)	9.78–78.26	42.52 (21.59)	11.23–81.70
Inter-rater: Video score							
Assessor B	15	54.20 (25.35)	7.61–92.39	44.42 (22.58)	6.52–78.26	47.48 (23.32)	7.07–82.97
Assessor D	15	52.36 (23.96)	8.70–91.30	42.50 (21.37)	7.61–78.26	45.55 (22.22)	8.15–82.97
Intra-rater: Video score							
Assessor B Live	10	42.71 (20.90)	13.04–81.52	34.13 (16.97)	9.78–66.30	36.97 (18.49)	11.41–71.01
Assessor B Video	10	45.22 (21.70)	7.61–82.61	37.39 (19.63)	6.52–70.65	39.26 (19.50)	7.07–71.38
Test–retest: (assessor A) Live score							
Test	22	47.04 (21.52)	11.96–85.87	38.14 (18.59)	9.78–78.26	41.02 (19.35)	11.23–79.35
Retest	22	46.89 (21.40)	11.96–88.04	38.65 (20.35)	8.70–85.87	41.46 (20.52)	11.41–85.51

ICC: intraclass correlation coefficient; CI: 95% confidence interval of ICC; SEM: standard error of measurement.

Table 4. Reliability estimates for the Danish *Challenge*.

		Best score total		First score total		Mean score total	
Testing scenario	*n*	ICC (95% CI)	SEM	ICC (95% CI)	SEM	ICC (95% CI)	SEM
(i) Inter-rater scenarios							
Live score (assessor A vs. B + C)	45	0.998 (0.998–0.999)	0.87	0.995 (0.991–0.997)	1.48	0.998 (0.996–0.999)	1.04
Live score (assessor A vs. B)	29	0.998 (0.996–0.999)	0.97	0.994 (0.987–0.997)	1.61	0.997 (0.995–0.999)	1.02
Live score (assessor A vs. C)	16	0.999 (0.998–0.999)	0.66	0.997 (0.991–0.999)	1.23	0.998 (0.995–0.999)	0.99
Video-Video score (assessor B vs. D)	15	0.991 (0.963–0.997)	2.24	0.981 (0.939–0.994)	3.00	0.988 (0.952–0.996)	2.41
(ii) Intra-rater scenarios							
Live-Video score (assessor B vs. B)	10	0.977 (0.895–0.994)	3.25	0.951 (0.780–0.988)	4.00	0.975 (0.889–0.994)	2.96
(iii) Test–retest scenario							
Live score (assessor A)	22	0.991 (0.979–0.996)	2.03	0.928 (0.836–0.969)	5.20	0.976 (0.943–0.989)	3.11
GMFCS level breakdown							
(iv) Inter-rater scenario							
GMFCS I – Live score (assessor A vs. B + C)	25	0.999 (0.998–0.999)	0.88	0.995 (0.988–0.998)	1.73	0.998 (0.995–0.999)	1.18
GMFCS II – Live score (assessor A vs. B + C)	20	0.996 (0.990–0.999)	0.88	0.993 (0.983–0.997)	1.09	0.997 (0.993–0.999)	0.71
GMFCS I – Video score (assessor B vs. D)	8	0.997 (0.955–0.999)	1.76	0.991 (0.959–0.998)	2.46	0.993 (0.958–0.999)	2.27
GMFCS II – Video score (assessor B vs. D)	7	0.950 (0.760–0.991)	2.69	0.899 (0.576–0.982)	3.57	0.943 (0.729–0.989)	2.68
(v) Intra-rater scenario: GMFCS I – Live-Video score (assessor B vs. B)	4	0.995 (0.922–0.999)	2.30	0.989 (0.845–0.999)	2.90	0.995 (0.925–0.999)	2.05
GMFCS II – Live-Video score (assessor B vs. B)	6	0.931 (0.288–0.991)	3.75	0.841 (0.046–0.977)	4.73	0.921 (0.216–0.989)	3.44
(vi) Test–retest scenario: GMFCS I – Live score (assessor A)	9	0.994 (0.974–0.998)	1.92	0.909 (0.650–0.979)	7.09	0.974 (0.889–0.994)	3.79
GMFCS II – Live score (assessor A)	13	0.982 (0.943–0.995)	2.09	0.938 (0.810–0.981)	3.33	0.969 (0.902–0.990)	2.52

ICC: intra class correlation coefficient; CI: 95% confidence interval of ICC; SEM: standard error of measurement; GMFCS: Gross Motor Function Classification System.

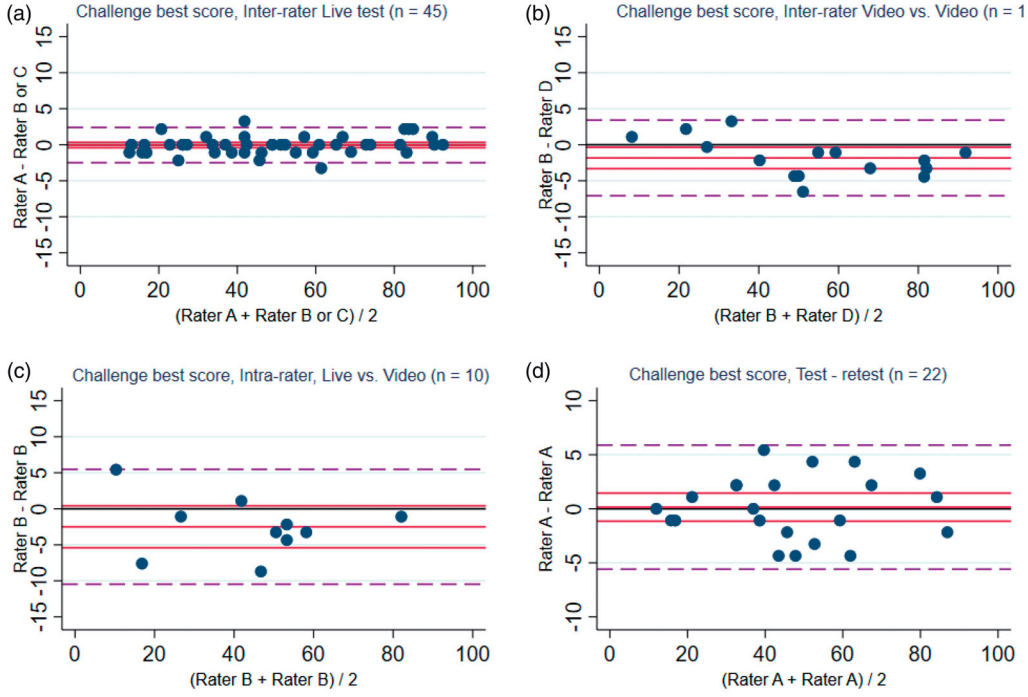

Figure 1. Agreement illustrated by the Bland–Altman plot with comparison between different pairings. The difference between two ratings on the vertical axis is plotted against the average of the two ratings on the horizontal axis. The middle horizontal line reflects the mean difference, and the upper and the lower line the limits of agreement. (a) Inter-rater live assessments. The middle horizontal line reflects the mean difference –0.046 and 95% CI (–0.420; 0.328) and the upper and the lower line the LOA (–2.536; 2.444). (b) Inter-rater video assessments. The middle horizontal line reflects the mean difference –1.839 and 95% CI (–3.322; –0.357) and the upper and the lower line the LOA (–7.194; 3.516). (c) Intra-rater live vs. video assessments. The middle horizontal line reflects the mean difference –2.5 and 95% CI (–5.410; 0.410) and the upper and the lower line the LOA (–10.637; 5.637). (d) Test–retest; two live assessments same day. The middle horizontal line reflects the mean difference 0.150 and 95% CI (–1.150.; 1.449) and the upper and the lower line the LOA (–5.712; 6.011).

A study using the newly created Brazilian-Portuguese 25-item *Challenge* version revealed excellent reliability with ICC estimates $>$ 0.95 and narrow confidence intervals between 0.96 and 1.00 for intra-rater and inter-rater reliability [27]. These findings are in line with the original reliability estimates [22]. Our findings are comparable to both studies and add to the emerging evidence picture of sound psychometric measurement properties of the *Challenge*.

The indication of the slightly greater disagreement in ratings (albeit small) in the mid-range of the scale for the intra-rater (live versus video, Assessor B) and test–retest (assessor A) may reflect challenges experienced when rating the accuracy component of a child's performance as the 40–50% total score range is consistent with having many items that scores of the "1" to "2" reflecting foot placement errors such as stepping on the track. These errors occur quickly and can be difficult to see as they can depend on assessor viewing angle if marginal steps on the line, and may be easier to spot on video.

While we achieved the target sample size of 45 for the inter-rater reliability analysis, we were only able to enroll 22 children

for test–retest reliability estimations (Table 4). The reason for not reaching the target of $n = 30$ was families limited time to stay for two test sessions on same day. A sample of 22 is still reasonable for reliability estimation based on other motor measures reliability studies in children with disabilities, as noted in the sample size section earlier.

For test–retest reliability evaluation, the decision on the time interval between the two assessments is, according to Terwee et al., not theoretical but instead relies on common sense meaning that it is crucial to consider the stability of participants' and assessors' characteristics between assessments as well as practical testing issues [41]. Our test–retest interval differed from the original *Challenge* study in which a 2–3 week retest interval was used to partially mimic week to week variability [22] and the Brazilian *Challenge* study where a 7–10 day interval was used [27]. In our study, we took a conservative approach to achieve the ultimate stability scenario, and used a same day retest interval. The same-day retesting was helpful for maximizing sample size due to families' availability. The obvious reliability impact consideration was the potential for inducement of a physical/mental fatigue effect. However, there was no direct evidence of fatigue since the results revealed no systematic differences in best score total between the two assessments (Figure 1(d)).

This study reported an MDC_{80} estimate of 3.7 points for the *Challenge*'s primary score (best score total) which is close to the MDC_{80} of 3.5 points in the original *Challenge* study from a one to three week retest interval [22], and less than an estimated MDC_{80} of 4.9 points for a 7–10 day retest interval (calculated here from data in Table 3 of that paper) in the Brazilian *Challenge* study [27]. In the original study, the researchers proposed that this result could be both possible and meaningful to achieve within a motor skills training program, e.g., gaining one point on 3–4 items [22]. Larger change requirements associated with the more traditional MDC_{90} and MDC_{95} may be a larger change than can be achieved within a single intervention [22]. As well, this higher level of confidence (smaller CI range) offered by the MDC_{90} and MDC_{95} for detecting a pre-/post-intervention difference may not be necessary for decisions related to motor skill change [18].

Since the *Challenge* is a newly developed measure, there is still need for intervention studies to determine what would be considered both clinically meaningful and achievable as far as change scores [22,24]. Results from a feasibility study with the *Challenge-20* with children in GMFCS level I and GMFCS level II aged 8–17 years ($n = 20$) showed significant mean gains of 2.8 points on the *Challenge-20* associated with a sports-based skills training program. These provide a first indication of the ability of the shorter version *Challenge-20* to detect change in advanced motor skill performance. From a concurrent validity perspective, these gains were accompanied by clinically important and significant changes of about four points in individualized physical activity-focused goals as measured by the Canadian Occupational Performance Measure [25].

One key aim of rehabilitation for children with CP is to maximize each individual's ability, keeping in mind their personal goals and expectations, to participate in everyday and recreational activities [42]. Optimization of engagement and motivation is important as part of a positive testing environment to support the child in demonstrating their best abilities and generally lead to a positive testing and subsequent goal setting experience. The *Challenge,* together with its engagement guidelines [30] appears well-suited given these promising reliability and MDC estimates to fill this role when physiotherapists are assessing advanced motor skills of children with CP in GMFCS levels I and II, and has strong potential to support both the individualized goal based planning and evaluation of sports-linked skills based interventions [42].

Limitations

The high ICC's in this study might have been influenced by several factors [41]. Children in this study's convenience sample were by chance heterogeneous in age, gender and GMFCS levels. As the results show, the total scores were spread across the scale for the best score total, varying from 11.96% to 92.39% (Table 3). Since an ICC reflects a measure's ability to discriminate among subjects, large inter-subject variance in a study sample has a tendency to inflate an ICC [41]. However, this scoring range was comparable to the original *Challenge* study and as such reflects one of the goals of the measure in its creation which was to discriminate among children within/across GFMCS levels I and II [22]. Another limitation might be that only one experienced physiotherapist gave the *Challenge* instructions to all children, which may have minimized the rater variance within the ICC's if compared to test instructions provided by different instructors.

The sample size was a limiting factor for interpreting the Bland–Altman plots. While there was a suggestion of rating bias in the scores on the 40–50% range of the *Challenge*, additional data (ideally a sample size of 45 for all comparisons) across the range of scores would have helped to delineate whether this was a reflection of rating challenges or was just a spurious pattern with the smaller data set. For the estimation of inter-rater reliability using video recording in a subset, we managed to record only $n = 15$ children (Table 4). This smaller sample reduced the ability to directly compare to live inter-rater results. This was due to the logistical challenge of availability of the needed extra personnel to do the video recordings.

The *Challenge* (20-item version) was used for all score calculations in this study [24]. This should be considered, when comparing the total scores results (Table 2) with the original *Challenge* study and those of the Brazilian study as both reported total score on the 25-item version. However, the *Challenge* (20-item version) aims to enhance sensitivity to change across the score range and to reduce the number of items of the *Challenge* by removing those items with poor discrimination or difficulty level overlap. The 20-item version is now the *Challenge* version to be used clinically as well as in research work (Personal communication; FV Wright, August 2020).

Conclusions

The *Challenge* revealed excellent inter-rater, intra-rater and test–retest reliability in both live-testing and video-recorded assessment scoring, after translation and cultural adaptation. For these results to apply to clinical physiotherapists, they need to practice and pass the *Challenge* criterion-based training and online calibration to become sufficiently competent in its use. Hereafter, physiotherapists can choose to score the *Challenge* from live- or video-recorded assessments.

The *Challenge* was developed both to help clinicians establish goals with children/families at the start of a block of physiotherapy, or prior to orthopedic surgery or spasticity reduction intervention, or upon entering a community-based physical activity program, and to evaluate outcome in relation to interventions designed to address advanced motor skill performance. Thus, its sensitivity and responsiveness to change after such interventions is also important to determine, and next stage research needs to

be conducted in this regard before it can confidently be used as an outcome tool.

Acknowledgements

Thanks, to the research team at Bloorview Research Institute Toronto, Canada for collaboration to fulfill our task, while the development of the original *Challenge*, was still going on. Thanks to all the children and their parents, who gave their time and contributed with great enthusiasm by participating into this study, it was fun. Thanks to Trine Friis Gehlert, Michael Overgaard, and Lone Nielsen for working with testing, to professional's participation into the translation processes and also Therese Koops Groenborg for help, advice, and guidance on statistics, who all together have contributed to get this work done. Funding of this study was with donations from the Danish Physiotherapist Research Foundation and from Health Research foundation Central Region Denmark. Authors have stated their disclosures.

Disclosure statement

One of the authors F. Virginia Wright is the lead developer of the original English version of the *Challenge* measure. This author do not have any personal financial interest in the measure, i.e., do not receive any money or other compensation from its sharing and use.

ORCID

Kirsten Nordbye-Nielsen http://orcid.org/0000-0003-3332-9630
Thomas Maribo http://orcid.org/0000-0003-0856-6837
F. Virginia Wright http://orcid.org/0000-0002-9713-4536

Data availability statement

The *Challenge* material in Danish is only available after training and passing the online criterion calibration. Contact F. Virginia Wright PT, PhD, Bloorview Research Institute for further information, vwright@hollandbloorview.ca, or Kirsten Nordbye-Nielsen, Aarhus University Hospital Denmark, kirsn1@rm.dk. This study was conducted at the Department of Children's Orthopedics, Aarhus University Hospital, Denmark.

References

[1] Bax M, Goldstein M, Rosenbaum P, et al. Proposed definition and classification of cerebral palsy, April 2005. Dev Med Child Neurol. 2005;47(8):571–576.
[2] Chiarello LA, Palisano RJ, Bartlett DJ, et al. A multivariate model of determinants of change in gross-motor abilities and engagement in self-care and play of young children with cerebral palsy. Phys Occup Ther Pediatr. 2011;31(2):150–168.
[3] van Gorp M, Van Wely L, Dallmeijer AJ, et al. Long-term course of difficulty in participation of individuals with cerebral palsy aged 16 to 34 years: a prospective cohort study. Dev Med Child Neurol. 2019;61(2):194–203.
[4] Graham HK, Rosenbaum P, Paneth N, et al. Cerebral palsy. Nat Rev Dis Primers. 2016;2:15082.
[5] Narayanan UG. Should children with cerebral palsy exercise? Dev Med Child Neurol. 2015;57(7):597–598.
[6] Novak I, Hines M, Goldsmith S, et al. Clinical prognostic messages from a systematic review on cerebral palsy. Pediatrics. 2012;130(5):e1285–e1312.
[7] Frøslev-Friis C, Dunkhase-Heinl U, Andersen JD, et al. Epidemiology of cerebral palsy in Southern Denmark. Dan Med J. 2015;62:A4990.
[8] Westbom L, Hagglund G, Nordmark E. Cerebral palsy in a total population of 4–11 year olds in southern Sweden. Prevalence and distribution according to different CP classification systems. BMC Pediatr. 2007;7:41.
[9] Palisano R, Rosenbaum P, Walter S, et al. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol. 1997;39(4):214–223.
[10] Rasmussen HM, Nordbye-Nielsen K, Moller-Madsen B, et al. The Danish cerebral palsy follow-up program. Clin Epidemiol. 2016;8:457–460.
[11] Alriksson-Schmidt AI, Arner M, Westbom L, et al. A combined surveillance program and quality register improves management of childhood disability. Disabil Rehabil. 2017;39(8):830–836.
[12] Oeffinger D, Bagley A, Rogers S, et al. Outcome tools used for ambulatory children with cerebral palsy: responsiveness and minimum clinically important differences. Dev Med Child Neurol. 2008;50(12):918–925.
[13] Clutterbuck GL, Auld ML, Johnston LM. High-level motor skills assessment for ambulant children with cerebral palsy: a systematic review and decision tree. Dev Med Child Neurol. 2020;62(6):693–699.
[14] Hanna SE, Rosenbaum PL, Bartlett DJ, et al. Stability and decline in gross motor function among children and youth with cerebral palsy aged 2 to 21 years. Dev Med Child Neurol. 2009;51(4):295–302.
[15] Russell D, Rosenbaum P, Wright M, et al. Gross motor function measure (GMFM-66 and GMFM-88) user's manual. London: McKeith Press; 2013.
[16] Smits DW, Gorter JW, Hanna SE, et al. Longitudinal development of gross motor function among Dutch children and young adults with cerebral palsy: an investigation of motor growth curves. Dev Med Child Neurol. 2013;55(4):378–384.
[17] Alotaibi M, Long T, Kennedy E, et al. The efficacy of GMFM-88 and GMFM-66 to detect changes in gross motor function in children with cerebral palsy (CP): a literature review. Disabil Rehabil. 2014;36(8):617–627.
[18] Himuro N, Abe H, Nishibu H, et al. Easy-to-use clinical measures of walking ability in children and adolescents with cerebral palsy: a systematic review. Disabil Rehabil. 2017;39(10):957–968.
[19] Rosenbaum PL, Russell DJ, Cadman DT, et al. Issues in measuring change in motor function in children with cerebral palsy: a special communication. Phys Ther. 1990;70(2):125–131.
[20] Novak I, McIntyre S, Morgan C, et al. A systematic review of interventions for children with cerebral palsy: state of the evidence. Dev Med Child Neurol. 2013;55(10):885–910.
[21] Wilson A, Kavanaugh A, Moher R, et al. Development and pilot testing of the challenge module: a proposed adjunct to the Gross Motor Function Measure for high-functioning children with cerebral palsy. Phys Occup Ther Pediatr. 2011;31(2):135–149.
[22] Wright FV, Lam CY, Mistry B, et al. Evaluation of the reliability of the challenge when used to measure advanced

motor skills of children with cerebral palsy. Phys Occup Ther Pediatr. 2018;38(4):382–394.
[23] Clutterbuck G, Auld M, Johnston L. Active exercise interventions improve gross motor function of ambulant/semi-ambulant children with cerebral palsy: a systematic review. Disabil Rehabil. 2019;41(10):1131–1151.
[24] Wright V, Avery L, Fehlings D, et al. Assessing advanced motor skills in young people with cerebral palsy in GMFCS levels I and II: Rasch analysis of the challenge. Dev Med Child Neurol. 2016;58:77–78.
[25] Hilderley AJ, Fehlings D, Chen JL, et al. Comparison of sports skills movement training to lower limb strength training for independently ambulatory children with cerebral palsy: a randomised feasibility trial. Disabil Rehabil. 2020;1–9.
[26] Levac D, McCormick A, Levin MF, et al. Active video gaming for children with cerebral palsy: does a clinic-based virtual reality component offer an additive benefit? A pilot study. Phys Occup Ther Pediatr. 2018;38(1):74–87.
[27] Sousa Junior RR, Gontijo APB, Santos TRT, et al. Measurement properties and translation to Brazilian-Portuguese of the challenge for children and adolescents with cerebral palsy. Phys Occup Ther Pediatr. 2020;1–18.
[28] World Health Organization. Process of translation and adaptation of instruments. Geneva (Switzerland): World Health Organization; 2011.
[29] Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000;25(24):3186–3191.
[30] Gibson BE, Mistry B, Wright FV. Development of child and family-centered engagement guidelines for clinical administration of the challenge to measure advanced gross motor skills: a qualitative study. Phys Occup Ther Pediatr. 2018; 38(4):417–426.
[31] Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002; 21(9):1331–1335.
[32] Tustin K, Gimeno H, Morton E, et al. Rater reliability and scoring duration of the Quality Function Measure in ambulant children with hyperkinetic movement disorders. Dev Med Child Neurol. 2016;58(8):822–828.
[33] Klingels K, De Cock P, Desloovere K, et al. Comparison of the Melbourne Assessment of Unilateral Upper Limb Function and the Quality of Upper Extremity Skills Test in hemiplegic CP. Dev Med Child Neurol. 2008;50(12):904–909.
[34] Thomas SS, Buckon CE, Phillips DS, et al. Interobserver reliability of the gross motor performance measure: preliminary results. Dev Med Child Neurol. 2001;43(2):97–102.
[35] Sorsdahl AB, Moe-Nilssen R, Strand LI. Observer reliability of the Gross Motor Performance Measure and the Quality of Upper Extremity Skills Test, based on video recordings. Dev Med Child Neurol. 2008;50(2):146–151.
[36] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–428.
[37] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–163.
[38] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–310.
[39] Wong RK, McEwan J, Finlayson D, et al. Reliability and validity of the Acquired Brain Injury Challenge Assessment (ABI-CA) in children. Brain Inj. 2014;28(13–14):1734–1743.
[40] Almeida KM, Albuquerque KA, Ferreira ML, et al. Reliability of the Brazilian Portuguese version of the Gross Motor Function Measure in children with cerebral palsy. Braz J Phys Ther. 2016;20(1):73–80.
[41] Terwee CK, Vet CWeV, Mokkin LB. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.
[42] Reedman S, Boyd RN, Sakzewski L. The efficacy of interventions to increase physical activity participation of children with cerebral palsy: a systematic review and meta-analysis. Dev Med Child Neurol. 2017;59(10):1011–1018.