



Reliability and construct validity of the Duchenne Video Assessment

Marielle G. Contesse PhD, MPH¹  | Amber T. L. Sapp PT, DPT¹ |
Susan D. Apkon MD² | Linda P. Lowes PT, PhD³  | Laura Dalle Pазze BA⁴ |
Mindy G. Leffler MEd¹

¹Casimir, Plymouth, Massachusetts

²Children's Hospital Colorado, Aurora, Colorado

³Nationwide Children's Hospital, Columbus, Ohio

⁴Charley's Fund, New York, New York

Correspondence

Marielle G. Contesse, Casimir, 36 Cordage Park Circle, S. 300, Plymouth, MA 02360, USA.
Email: mariellec@casimirtrials.com

Funding information

Charley's Fund, Grant/Award Number: N/A

Abstract

Introduction: The Duchenne Video Assessment (DVA) assesses quality of movement as an indication of Duchenne muscular dystrophy (DMD) disease severity. Caregivers video record patients performing home-based movement tasks using a mobile application, and physical therapists (PTs) rate the videos using scorecards with prespecified compensatory movement criteria. Reliability and construct validity of the DVA were tested using video and Pediatric Outcomes Data Collection Instrument (PODCI) data from patients with DMD and healthy controls from a separate study.

Methods: Fifteen PTs were trained and certified as DVA raters. All raters scored videos of five subjects performing each movement task; nine raters rescored the same videos four weeks later. Three raters scored videos from an average of 25 subjects for each movement task. Aggregate scores were used to test construct validity. An expert DMD clinician assigned each video to a severity group for known-groups analyses. Differences between rater scores across severity groups were tested and correlations between DVA and PODCI scores were calculated.

Results: Inter-rater reliability (intraclass correlation coefficient [ICC]) between all 15 raters ranged from 0.70 to 0.97 for all movement tasks. Mean intra-rater reliability ICC for nine raters ranged from 0.82 to 0.98 for all movement tasks. There were statistically significant differences between known severity groups for all movement tasks. The DVA correlated strongly with related PODCI constructs of physical function and weakly with unrelated constructs.

Discussion: The DVA was found to be a reliable and valid tool for measuring quality of movement as an indication of disease severity.

Abbreviations: CI, confidence interval; DMD, Duchenne muscular dystrophy; DVA, Duchenne Video Assessment; EK, Egen Klassifikation; ICC, intraclass correlation coefficient; IQR, interquartile range; NSAA, North Star Ambulatory Assessment; PODCI, Pediatric Outcomes Data Collection Instrument; PT, physical therapist; PUL, Performance of Upper Limb.

Conferences

A related poster was presented at the virtual World Muscle Society Conference from September 28 to October 2, 2020, and a related presentation was given at the virtual EveryLife Foundation Rare Disease Scientific Workshop on December 15, 2020.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 Casimir LLC. *Muscle & Nerve* published by Wiley Periodicals LLC.

KEYWORDS

clinical outcome assessment, Duchenne muscular dystrophy, eHealth, reliability, validity

1 | INTRODUCTION

Current Duchenne muscular dystrophy (DMD) clinical trials use a combination of timed function tests and clinician-rated assessments of movement to determine the efficacy of potential therapeutics.^{1,2} Timed function tests are typically only appropriate for a small subset of the DMD population,¹⁻³ leading clinical trials to require narrow inclusion criteria to be able to detect changes in function.⁴ Early in the disease progression, timed function tests have failed to demonstrate responsiveness to changes in younger patients with DMD.⁵ Non-ambulatory patients with DMD are not able to perform timed function tests, such as the 6-min walk or 4-stair climb, and would be excluded from trials using those outcome measures to determine efficacy.^{2,3} Using assessment tools that can encompass a wider DMD population may not only improve the ability of clinical trials to recruit subjects but also enable access to data assessing efficacy across the entire range of participants that may benefit from a potential therapeutic.

Evaluating the quality rather than the speed of movement allows for the quantification of predictable compensatory movement factors that coincide with progression of muscle weakness across the entire DMD population. People compensate for muscle weakness by changing their movement strategy.⁶⁻⁹ Many clinician-rated assessments of movement differentiate between uncompensated movement, compensated movement, and inability to perform a task.^{10,11} They do not delineate between different levels of compensated movement for each task, which may limit their ability to detect finer functional changes over the short term that are clinically meaningful.¹² Outcome measures are needed that can detect changes in a short time frame to reduce patient participation time in a clinical trial placebo group given the rapid rate of decline in physical strength and function of individuals with DMD.²

The Duchenne Video Assessment (DVA) addresses the need for DMD outcome measures that can detect functional changes over the short term and be applied to the entire population. Rather than measuring best performance in a clinic setting, the DVA measures habitual performance in the home environment. Using a secure mobile application, caregivers record videos of patients doing specific movement tasks at home. Instead of containing test items that are purely used to assess function (eg, stacking cans, hop on one leg),^{10,11} the DVA only includes movement tasks that are either activities of daily living (eg, putting a t-shirt on, eating) or foundational tasks of daily life (eg, walking, climbing stairs). DVA-certified physical therapists (PTs) assess the compensations used to perform each task using scorecards. Previous research has established the validity of movement task selection and scorecards.¹³⁻¹⁵ This study proceeded to evaluate the reliability and construct validity of the DVA.

2 | METHODS

2.1 | Duchenne Video Project – Video data collection

2.1.1 | Study design and subjects

In a longitudinal study (Duchenne Video Project) of male participants with and without DMD conducted from August 2018 to March 2020, caregivers collected video data of participants performing specific DVA movement activities over time for future outcome measure testing. Participants were recruited from across the United States by Casimir, Parent Project Muscular Dystrophy, Charley's Fund, and Jett Foundation through social media and email outreach.

Participants were recruited into two age cohorts, ≥ 7 y (Cohort 1) and 4–6 y (Cohort 2). Inclusion criteria for all subjects included ability to follow movement instructions, proficiency in English, and access to a smartphone. Participants enrolled into a DMD subgroup were required to have a confirmed diagnosis and had to have sufficient movement skills to perform tasks like pick up coins or hold a pen. Cohort 1 consisted of six subgroups: (1) Early Ambulatory DMD, (2) Late Ambulatory DMD, (3) Non-DMD Ages 7–11, (4) Early Non-Ambulatory DMD, (5) Late Non-Ambulatory DMD, and (6) Non-DMD Ages 12–16. Cohort 2 consisted of two subgroups: (1) Young DMD Ages 4–6 and (2) Non-DMD Ages 4–6. Cohort 1 completed video assessments at baseline, 30 wk post-baseline, and 42 wk post-baseline. Cohort 2 completed video assessments at baseline and 12 wk post-baseline.

The subgroup determination was based on caregiver report. Participants in both ambulatory groups did not need to use a wheelchair to move approximately 10 m and were subdivided by the ability to rise from the floor within 30 s without assistance or use of furniture. Those that could were considered early ambulatory, and those that could not were classified as late ambulatory. Participants in both non-ambulatory groups needed to use a wheelchair to move approximately 10 m and were subdivided based on arm function. Those in the early non-ambulatory group could still bring their hand to their mouth and those in the late non-ambulatory group had lost that ability but could still use their hands to do tasks such as pick up coins or hold a pen.

The study received ethical approval from Quorum Review, a central Institutional Review Board. All minor participants provided informed assent after their parents/legal guardians provided informed consent, and all adult participants provided informed consent.

2.1.2 | Measures

Caregivers were provided with a training manual, online training videos, and study supplies. The training manual described how to

register for, download, and use the study mobile application to submit videos securely. It also described how to standardize the set-up, lighting, clothing, surfaces, and instructions during video capture. Caregivers were invited to record their own videos in the mobile application after submitting a signed training documentation form. Of note, the Duchenne Video Project provided links to training videos that caregivers could access online, but commercial deployments of the DVA include training videos and documentation built into the study mobile application.

The movement activities were assigned to each subgroup based on age and functional status, and subjects could skip an assigned movement activity if they were not able to perform the task. Study staff monitored data collection to ensure that each video met quality standards. If a video did not meet quality standards, caregivers were asked to re-record the video. Caregivers completed the Pediatric Outcomes Data Collection Instrument (PODCI) questionnaire before or after the video assessment at each timepoint within the same 2-wk data collection window. The PODCI includes six core scales: (1) Upper Extremity and Physical Function, (2) Transfer and Basic Mobility, (3) Sports and Physical Functioning, (4) Pain/Comfort, (5) Happiness, and (6) Global Functioning. The Global Functioning scale is a combination of the Upper Extremity and Physical Function, Transfer and Basic Mobility, Sports and Physical Functioning, and Pain/Comfort scales.

The PODCI score values range from 0 (poor outcome/worse health) to 100 (best possible outcome/best health).

The 30-wk Cohort 1 data and baseline Cohort 2 data were used during the testing of the reliability and construct validity of the DVA and will be referred to as the test data set. Table 1 describes the movement tasks that were assigned to each subgroup and the number of participants within each subgroup that completed each movement task in the test data set.

2.2 | Reliability and construct validity testing

2.2.1 | Study design and raters

Reliability and construct validity testing were conducted with PTs from October 2019 to January 2020. Both PTs who specialize and who do not specialize in DMD were included in the testing to determine the level of DMD expertise required to be a reliable DVA rater. DMD specialist PTs who had assessed at least 50 patients with DMD were recruited from a list of United States PTs provided by a DMD PT key opinion leader (L.L.). The non-DMD specialist PTs were recruited through convenience sampling in the Seattle geographic area.

TABLE 1 Movement tasks assigned to each subgroup in the test data set

Movement task	Subgroups							
	A N = 9	B N = 8	C N = 5	D N = 10	E N = 10	F N = 4	G N = 11	H N = 6
Climb 5 stairs	X	X ^a	X				X	X
Run	X		X				X ^b	X
Walk	X	X ^c	X				X	X
Jump forward	X ^b		X				X ^b	X
Sit up	X	X	X				X	X
Stand up from sitting on floor	X		X				X	X
Stand up from supine	X		X				X	X
Stand up from sitting on couch	X	X ^b	X					
Raise hands above head		X ^d		X		X		
Roll over in bed		X		X		X ^d		
Take t-shirt off		X ^d		X		X		
Put t-shirt on		X ^d		X		X		
Shift weight in bed				X	X ^c	X		
Eat 10 bites				X	X ^b	X		
Put arm on armrest – Right arm				X	X	X		
Put arm on armrest – Left arm				X	X	X		
Reach across table to grab a cell phone				X	X ^b	X		

Note: Group A: Early Ambulatory DMD; Group B: Late Ambulatory DMD; Group C: Non-DMD Ages 7–11; Group D: Early Non-Ambulatory DMD; Group E: Late Non-Ambulatory DMD; Group F: Non-DMD Ages 12–16; Group G: Young DMD Ages 4–6; Group H: Non-DMD Ages 4–6.

^aMissing data for five subjects who could no longer perform this task.

^bMissing data for one subject who could no longer perform this task.

^cMissing data for two subjects who could no longer perform this task.

^dMissing data for one subject who did not submit videos for this task.

Fifteen PT raters with an active license and a minimum of 1 y of experience completed the DVA Rater Training Program prior to scoring videos. After watching training videos, the PTs completed a certification test that consisted of scoring one video for each movement task in an online scoring dashboard. Raters had to pass each scorecard with at least 80% accuracy to become a certified rater, and they could take the certification test up to three times total.

Following certification, 15 raters (seven DMD specialist and eight non-DMD specialist) scored videos of a different set of five participants performing each movement task for inter-rater reliability testing. Nine raters (four DMD specialist and five non-DMD specialist) re-scored the same videos at least 4 wk later for intra-rater reliability testing. Three raters scored the remainder of the videos in the test data set, which consisted of an average of 25 subjects for each movement task, for construct validity testing.

An expert DMD clinician (S.A.) created task-specific severity groups based on expert judgement and knowledge of the disease. Each video in the test data set was classified into a severity group based solely on the task being evaluated not on assigned global functional cohort to achieve greater specificity to the targeted muscle groups for each task. The severity groups included: (1) no weakness, (2) mild weakness, (3) moderate weakness, (4) severe weakness, and (5) cannot complete task. Group 5 was only used when a caregiver submitted a video of a DMD patient unable to complete the movement task.

2.2.2 | DVA severity percentage

Each movement task has a unique list of clinically meaningful compensations detailed in each scorecard.¹⁵ An example scorecard is provided in Figure S1, which is available online, with a description of the way the scorecard is used.

2.2.3 | Statistical analyses

Inter-rater reliability intraclass correlation coefficients (ICCs) were calculated for each movement task using two-way random-effects models (ICC (2, 1)),¹⁶ and the absolute agreement estimates were presented as ICCs and 95% confidence intervals (CIs). All reliability analyses were stratified by type of rater: DMD specialists and non-DMD specialists. The interpretation of the reliability is as follows: Poor <0.50, Moderate 0.50–0.74, Good 0.75–0.90, and Excellent >0.90.¹⁷ When comparing DMD specialists to non-DMD specialists, we identified any tasks that had an ICC difference of ≥ 0.15 , a threshold set by the study team to indicate meaningful differences in reliability.

Intra-rater reliability ICCs for each movement task were calculated by individual rater using two-way mixed-effects models (ICC (3, 1)),¹⁶ and the mean intra-rater reliability was calculated by movement task across all raters. The mean ICC and the minimum

TABLE 2 Inter-rater reliability: Absolute agreement by movement task (five subjects per task)

Movement task	DMD specialist physical therapists N = 7 ICC (95% CI) ^a	Non-DMD specialist physical therapists N = 8 ICC (95% CI) ^a	All physical therapists N = 15 ICC (95% CI) ^a
Climb 5 stairs	0.95 (0.85, 0.99)	0.94 (0.84, 0.99)	0.95 (0.86, 0.99)
Run	0.84 (0.59, 0.98)	0.77 (0.49, 0.97)	0.79 (0.55, 0.97)
Walk	0.79 (0.50, 0.97)	0.60 (0.27, 0.93)	0.70 (0.41, 0.95)
Jump forward	0.95 (0.85, 0.99)	0.96 (0.89, 1.00)	0.93 (0.82, 0.99)
Sit up	0.99 (0.97, 1.00)	0.95 (0.86, 0.99)	0.97 (0.91, 1.00)
Stand up from sitting on floor	0.97 (0.90, 1.00)	0.96 (0.88, 1.00)	0.96 (0.89, 1.00)
Stand up from supine	0.97 (0.90, 1.00)	0.93 (0.79, 0.99)	0.94 (0.84, 0.99)
Stand up from sitting on couch	0.95 (0.86, 0.99)	0.96 (0.89, 1.00)	0.96 (0.89, 0.99)
Raise hands above head	0.88 (0.69, 0.98)	0.93 (0.81, 0.99)	0.92 (0.78, 0.99)
Roll over in bed	0.98 (0.93, 1.00)	0.96 (0.87, 0.99)	0.96 (0.89, 0.99)
Shift weight in bed	0.92 (0.76, 0.99)	0.89 (0.71, 0.99)	0.91 (0.77, 0.99)
Take t-shirt off	0.81 (0.54, 0.97)	0.70 (0.38, 0.95)	0.76 (0.51, 0.96)
Put t-shirt on	0.89 (0.70, 0.99)	0.87 (0.66, 0.98)	0.87 (0.69, 0.98)
Eat 10 bites	0.93 (0.81, 0.99)	0.85 (0.62, 0.98)	0.89 (0.73, 0.99)
Put arm on armrest – Right arm	0.92 (0.77, 0.99)	0.91 (0.75, 0.99)	0.92 (0.79, 0.99)
Put arm on armrest – Left arm	0.90 (0.72, 0.99)	0.92 (0.78, 0.99)	0.91 (0.78, 0.99)
Reach across table to grab a cell phone	0.91 (0.76, 0.99)	0.89 (0.71, 0.99)	0.90 (0.76, 0.99)

^aThe ICC interpretation is as follows: poor <0.50, moderate 0.50–0.74, good 0.75–0.90, and excellent >0.90.

TABLE 3 Mean intra-rater reliability, by movement task (absolute agreement; five subjects per task, two timepoints)

Movement task	DMD specialist physical therapists N = 4 Mean ICC (min, max) ^a	Non-DMD specialist physical therapists N = 5 Mean ICC (min, max) ^a	All physical therapists N = 9 Mean ICC (min, max) ^a
Climb 5 stairs	0.90 (0.78, 0.97)	0.96 (0.94, 0.99)	0.93 (0.78, 0.99)
Run	0.94 (0.87, 1.00)	0.81 (0.71, 0.90)	0.87 (0.71, 1.00)
Walk	0.96 (0.93, 0.98)	0.70 (0.36, 0.98)	0.82 (0.36, 0.98)
Jump forward	0.99 (0.98, 0.99)	0.92 (0.81, 0.98)	0.95 (0.81, 0.99)
Sit up	0.97 (0.89, 0.99)	0.98 (0.94, 0.99)	0.97 (0.89, 0.99)
Stand up from sitting on floor	0.98 (0.97, 1.00)	0.98 (0.95, 0.99)	0.98 (0.95, 1.00)
Stand up from supine	0.94 (0.86, 0.98)	0.96 (0.86, 0.99)	0.95 (0.86, 0.99)
Stand up from sitting on couch	0.98 (0.96, 1.00)	0.97 (0.94, 1.00)	0.98 (0.94, 1.00)
Raise hands above head	0.95 (0.92, 0.97)	0.96 (0.89, 0.99)	0.96 (0.89, 0.99)
Roll over in bed	0.96 (0.95, 0.96)	0.96 (0.92, 0.99)	0.96 (0.92, 0.99)
Shift weight in bed	0.91 (0.89, 0.98)	0.96 (0.89, 1.00)	0.94 (0.89, 1.00)
Take t-shirt off	0.94 (0.87, 0.97)	0.89 (0.79, 1.00)	0.91 (0.79, 1.00)
Put t-shirt on	0.88 (0.85, 0.89)	0.90 (0.84, 0.95)	0.89 (0.84, 0.95)
Eat 10 bites	0.98 (0.94, 1.00)	0.93 (0.87, 0.97)	0.95 (0.87, 1.00)
Put arm on armrest – Right arm	0.98 (0.93, 1.00)	0.96 (0.92, 1.00)	0.97 (0.92, 1.00)
Put arm on armrest – Left arm	0.97 (0.94, 1.00)	0.97 (0.95, 1.00)	0.97 (0.94, 1.00)
Reach across table to grab a cell phone	0.74 (0.16, 0.97)	0.94 (0.87, 0.97)	0.85 (0.16, 0.97)

Abbreviations: Max, maximum; Min, minimum.

^aThe ICC interpretation is as follows: poor <0.50, moderate 0.50–0.74, good 0.75–0.90, and excellent >0.90.

and maximum ICC for the raters for each movement task are presented.

In the known-groups analyses, the aggregate severity percentages were compared to the clinician severity classification for all participants for each task. The Kruskal-Wallis test was used to determine whether there were significant differences in mean rank severity percentages across the severity groups, and the mean ranks, chi-squared test statistic, and *p*-values for each test were reported. This manuscript reports the results using the median criterion-level severity percentage of the three raters as the aggregate score, but additional analyses were conducted to evaluate the use of individual rater severity percentages and their mean, median, and maximum as aggregate scores. The median-level criterion severity percentage uses the median severity level selected within each compensatory criterion for the three raters and calculates a new score based on the sum of the severity points. This score allows a majority to rule on each compensatory criterion or selects the middle level if there are three different levels selected.

To evaluate convergent and discriminant validity, Spearman rank order correlation was used to calculate the degree of association (Spearman's rho) between DVA rater severity percentages and standardized PODCI scores for each movement task. Very strong ($|\rho| \geq 0.80$) or moderate ($|\rho| 0.60$ – 0.79) associations demonstrated convergent validity, and fair ($|\rho| 0.30$ – 0.59) or poor ($|\rho| < 0.30$) associations demonstrated discriminant validity.^{18,19} For this analysis, the median-level criterion severity percentage was used as the aggregate of the three

rater scores. A higher score on the PODCI indicates better health, and a lower DVA severity percentage indicates better health. Since the Upper Extremity and Physical Function, Transfer and Basic Mobility, and Sports and Physical Functioning scales relate to quality of movement, at least moderate correlation ($|\rho| \geq 0.60$) with the DVA rater severity percentages was expected. Since the Pain/Comfort and Happiness scales do not relate to quality of movement, fair or poor association ($|\rho| < 0.60$) with the DVA rater severity percentages was expected.

All analyses were conducted using Stata 16.0 (College Station, TX).

3 | RESULTS

3.1 | Duchenne Video Project participant and video information

Of the 63 Duchenne Video Project participants in the test data set, the median age was 11 y (interquartile range [IQR] 6, 15) and there were 15 (24%) non-DMD, 11 (17%) young DMD, 9 (14%) early ambulatory DMD, 8 (13%) late ambulatory DMD, 10 (16%) early non-ambulatory DMD, and 10 (16%) late non-ambulatory DMD participants. The participants were assigned movement tasks based on their subgroup, and Table 1 provides the number of participants that completed each movement task. Of the 498 videos submitted by caregivers in the test data set,

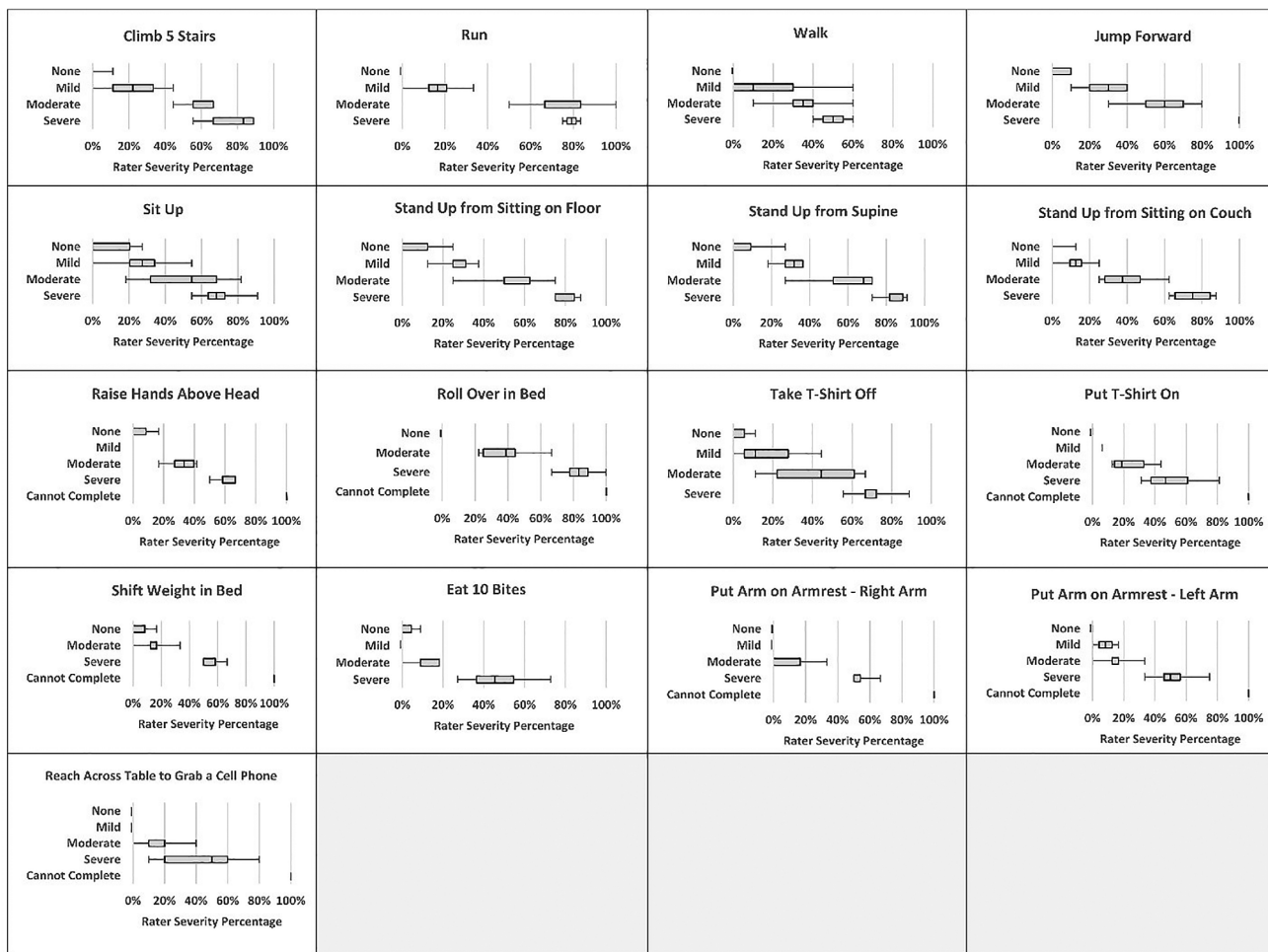


FIGURE 1 Box plots of aggregate rater severity percentage, by clinician-assigned severity group. The clinician-assigned severity groups are as follows: no weakness (“none”), mild weakness (“mild”), moderate weakness (“moderate”), severe weakness (“severe”), and cannot perform task (“cannot complete”)

65 (13%) did not meet quality standards and needed to be re-recorded.

3.2 | Rater demographic information

The characteristics of the PT raters are provided in Table S1.

3.3 | Inter-rater reliability

The inter-rater reliability between all 15 raters was excellent for 11 (65%), good for 5 (29%), and moderate for 1 (6%) of the movement tasks (Table 2). Between all 15 raters, the ICC for all movement tasks ranged from 0.70 to 0.97. The DMD specialist PTs had similar inter-rater reliability to the non-DMD specialist PTs for each movement task, except for the Walk movement task in which DMD specialist PTs had good reliability and non-DMD specialist PTs had moderate reliability.

3.4 | Intra-rater reliability

The mean intra-rater reliability for all nine raters was excellent for 13 (76%) and good for four (24%) of the movement tasks (Table 3). Among all nine raters, the mean ICC ranged from 0.82 to 0.98 for all movement tasks. The DMD specialist PTs had similar mean intra-rater reliability to the non-DMD specialist PTs for each movement task, except for the Walk movement task in which DMD specialist PTs had much higher mean intra-rater reliability and Reach Across Table to Grab a Cell Phone in which non-DMD specialist PTs had much higher mean intra-rater reliability.

3.5 | Known groups validity

For all 17 movement tasks, there were statistically significant differences ($p < .05$) in mean rank severity percentage between the severity groups (Table S2). The mean ranks for each group increased as the severity increased for 13 movement tasks, and there were four movement tasks (Raise Hands Above Head, Eat 10 Bites, Put Arm on

TABLE 4 Spearman rank correlation between Duchenne Video Assessment severity percentages and Pediatric Outcomes Data Collection Instrument scale scores

Movement task	Upper Extremity and Physical Function scale	Transfer and Basic Mobility scale	Sports and Physical Functioning scale	Pain/Comfort scale	Happiness scale	Global Functioning scale
Climb 5 stairs (N = 34) ^a	-0.70	-0.86	-0.82	-0.39	-0.54	-0.82
Run (N = 30) ^a	-0.46	-0.78	-0.80	-0.30	-0.63	-0.71
Walk (N = 36) ^a	-0.37	-0.61	-0.70	-0.29	-0.72	-0.61
Jump forward (N = 29) ^a	-0.61	-0.81	-0.80	-0.27	-0.55	-0.76
Sit up (N = 38) ^a	-0.71	-0.77	-0.75	-0.53	-0.46	-0.81
Stand up from sitting on floor (N = 31) ^a	-0.69	-0.86	-0.79	-0.31	-0.56	-0.79
Stand up from supine (N = 31) ^a	-0.64	-0.89	-0.80	-0.40	-0.57	-0.83
Stand up from sitting on couch (N = 20)	-0.72	-0.87	-0.82	-0.55	-0.39	-0.83
Raise hands above head (N = 18)	-0.32	-0.75	-0.54	0.05	-0.39	-0.59
Roll over in bed (N = 17)	-0.15	-0.78	-0.54	0.04	-0.23	-0.58
Shift weight in bed (N = 18)	-0.80	-0.73	-0.64	-0.52	-0.52	-0.78
Take t-shirt off (N = 18)	-0.03	-0.56	-0.30	-0.10	-0.41	-0.39
Put t-shirt on (N = 18)	-0.15	-0.63	-0.30	0.07	-0.21	-0.28
Eat 10 bites (N = 19)	-0.75	-0.88	-0.55	-0.55	-0.52	-0.77
Put arm on armrest - Right arm (N = 20)	-0.83	-0.94	-0.67	-0.55	-0.36	-0.79
Put arm on armrest - Left arm (N = 20)	-0.80	-0.88	-0.56	-0.46	-0.49	-0.74
Reach across table to grab a cell phone (N = 19)	-0.91	-0.87	-0.70	-0.55	-0.55	-0.89

Note: Spearman's rho interpretation is as follows: $|\geq 0.80|$ very strong, $|0.60-0.79|$ moderate, $|0.30-0.59|$ fair, $< 0.30|$ poor.

^aMissing data for one participant for Transfer and Basic Mobility, Happiness, and Global Functioning scales.

Armrest - Right Arm, and Reach Across the Table to Grab a Cell Phone) that did not have higher mean rank for the mild weakness group than for the no weakness group. For all 17 movement tasks, the results were similar when using the individual rater severity percentages and the median criterion-level, mean, median, and maximum severity percentages as the aggregate scores for the three raters.

Box plots of the aggregate rater severity percentages are presented in Figure 1, and they show a pattern of higher severity percentage IQRs with increasing severity group. The movement tasks related to elbow and wrist function (Eat 10 Bites, Putting Arms on Armrest, and Reach Across Table to Grab a Cell Phone) have more distinct severity percentage IQRs for moderate and severe weakness than for no and mild weakness.

3.6 | Convergent and discriminant validity

The correlation between the DVA severity percentages and the PODCI scales are presented in detail in Table 4. The Upper Extremity and Physical Function scale correlated most strongly with the DVA tasks related to elbow, wrist, and trunk function. The movement tasks that had very strong correlation with this scale were Reach Across the Table to Grab a Cell Phone, Put Arms on Armrest - Right Arm, Put Arms on Armrest - Left Arm, and Shift Weight in Bed. The Transfer and Basic Mobility scale correlated strongly with the majority of DVA tasks. The Sports and Physical Functioning scale correlated most strongly with the tasks related to lower body function. The movement tasks that had very strong correlation with this scale were Climb 5 Stairs, Stand Up from Sitting on Couch, Run, Jump Forward, and Stand Up from Supine. The Pain/Comfort and Happiness scales had poor or fair correlation with most DVA tasks.

4 | DISCUSSION

The DVA was found to be a reliable and valid tool for measuring quality of movement as an indication of disease severity. Both DVA inter-rater and intra-rater reliability were high overall, and reliability was similar between DMD specialist and non-DMD specialist PTs. Scoring reliability may be improved for movement tasks with lower reliability through changes to the DVA Rater Training Program that include more examples of the compensatory movements being scored. Based on the threshold of an ICC difference of ≥ 0.15 , the results suggest that additional training should be provided on the Walk task for non-DMD specialist PTs and provided on the Reach Across Table to Grab a Cell Phone task for the DMD specialist PTs to improve reliability.

The inter-rater reliability of the DVA is in line with existing clinician-rated assessments of movement compensations and exceeds the reliability of individual test items on the North Star Ambulatory Assessment (NSAA) and Egen Klassifikation (EK) Scale. The NSAA inter-rater reliability test included 17 evaluators rating three subjects, and the ICC ranged from 0.00 to 0.92 for the individual test items.¹⁰ The EK Scale inter-rater reliability test included 17 evaluators rating six subjects, and the weighted kappa ranged from 0.24 to 0.94 for the individual categories

of the assessment.²⁰ While the individual items on the NSAA and EK Scale did not all achieve high reliability, their total scores achieved high reliability with ICCs of 0.93 and 0.98, respectively. The DVA provides a composite score of individual compensatory movement criteria scores for each individual task, and the NSAA and EK Scale provide a composite score of individual task scores. Creating a composite score at the movement task level may allow the DVA to reach high reliability at the task level that the NSAA and EK Scale reach by creating a composite score of individual tasks. The Brooke Upper Extremity scale and Vignos Lower Extremity scale inter-rater reliability estimates were tested using four evaluators for 21 patients, and both scales achieved high reliability with ICCs of 0.87 and 0.96, respectively.²¹ The Performance of Upper Limb (PUL) inter-rater reliability was tested with 14 evaluators rating three patients, and it achieved an ICC of 0.96.²²

The DVA intra-rater reliability was excellent or good for all movement tasks, and it is comparable with other clinician-rated assessments of movement compensations. NSAA intra-rater reliability was tested with five evaluators rating one subject with 1 mo in between, and it achieved a percent agreement that ranged from 0.60 to 1.00 for all test items.¹⁰ The EK Scale intra-rater reliability was tested with seven evaluators rating six subjects with 6–8 wk in between, and it achieved an ICC of 0.98.²⁰ The Brooke Upper Extremity and Vignos Lower Extremity scales intra-rater reliability estimates were tested using four evaluators for 21 patients, and they achieved ICCs of 0.92 and 0.99, respectively.²¹ The PUL intra-rater reliability was tested with three evaluators assessing six patients with 1 h to 1 wk in between, and it achieved 100% agreement between the two timepoints.²²

Convergent validity has been established with the PODCI Global Functioning, Upper Extremity and Physical Function, Transfer and Basic Mobility, and Sports and Physical Functioning scales. Discriminant validity has been established with the Pain/Comfort and Happiness scales. Similarly, other studies have found that the PODCI domains most relevant to function and sensitive to differences between DMD functional groups are the Transfer and Basic Mobility, Upper Extremity and Physical Function, and Sports and Physical Functioning scales and that the Happiness and Pain/Comfort scales are not sensitive to differences between DMD functional groups.^{23,24}

The DVA is a clinical outcome assessment intended for use as an efficacy endpoint in clinical trials. There are not currently any clinical outcome assessments or biomarkers for DMD that are qualified by regulatory agencies for use as primary endpoints in clinical trials.^{25–28} Surrogate endpoints, such as muscle dystrophin expression, must be shown to predict or correlate with clinical benefit, often demonstrated through functional outcome measures.²⁹ Existing functional endpoints for DMD measure best performance in a clinical setting, which can be influenced by patient effort and encouragement by medical staff or caregivers.²⁵ The DVA provides an opportunity for clinical trials to evaluate the functional impact of a potential therapeutic on the daily lives of participants in their home environment.

This study has limitations. First, there were small sample sizes within some of the severity groups in the known-groups analyses, which prevented pairwise comparisons between severity groups. Second, since the expert DMD clinician saw videos of each participant performing multiple tasks, it is possible that assessment of participants

for individual tasks could have been influenced by an impression based on prior tasks. Third, these data are cross-sectional and did not allow for evaluation of associations between changes in function over time and changes in DVA scores.

This study demonstrated that the DVA is a valid tool that measures fine levels of disease severity while maintaining a high level of rater reliability. In addition to being used as a clinical trial endpoint, the DVA could be used for patient functional monitoring in a clinical setting and to support payer reimbursement decisions for medical and rehabilitation services. Future research will evaluate whether the DVA is able to detect functional change over a shorter duration than existing measures to address the current measurement challenges facing DMD clinical trials.

ACKNOWLEDGMENTS

The authors are grateful to the Duchenne Video Project families who contributed videos for testing and training. The authors thank the PTs who made this study possible, including Mark Bouma, Carla Corrado, Michael Kiefer, Melissa McIntyre, Natalie Miller, Karen Patterson, and Claudia Senesac. This research was funded by Charley's Fund.

ETHICAL PUBLICATION STATEMENT

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

CONFLICT OF INTEREST

None of the authors has any conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

Data available on request due to privacy/ethical restrictions

ORCID

Marielle G. Contesse  <https://orcid.org/0000-0002-4769-1617>

Linda P. Lowes  <https://orcid.org/0000-0003-4206-0557>

REFERENCES

- Domingos J, Muntoni F. Outcome measures in Duchenne muscular dystrophy: sensitivity to change, clinical meaningfulness, and implications for clinical trials. *Dev Med Child Neurol*. 2018;60:117-117. <https://doi.org/10.1111/dmcn.13634>
- Straub V, Mercuri E, Aartsma-Rus A, et al. Report on the workshop: meaningful outcome measures for Duchenne muscular dystrophy, London, UK, 30-31 January 2017. *Neuromuscul Disord*. 2018;28:690-701. <https://doi.org/10.1016/j.nmd.2018.05.013>
- Muntoni F. The development of antisense oligonucleotide therapies for Duchenne muscular dystrophy: report on a TREAT-NMD workshop hosted by the European Medicines Agency (EMA), on September 25th 2009. *Neuromuscul Disord*. 2010;20:355-362. <https://doi.org/10.1016/j.nmd.2010.03.005>
- McDonald C, Sweeney HL, Luo X, et al. Use of the six-minute walk distance (6MWD) across Duchenne muscular dystrophy (DMD) studies (P3.121). *Neurology*. 2016;86. https://n.neurology.org/content/86/16_Supplement/P3.121
- Arora H, Willcocks RJ, Lott DJ, et al. Longitudinal timed function tests in Duchenne muscular dystrophy: imaging DMD cohort natural history. *Muscle Nerve*. 2018;58:631-638. <https://doi.org/10.1002/mus.26161>
- de Souza CR, Betelli MT, Takazono PS, et al. Evaluation of balance recovery stability from unpredictable perturbations through the compensatory arm and leg movements (CALM) scale. *PLoS One*. 2019;14:e0221398. <https://doi.org/10.1371/journal.pone.0221398>
- Maki BE, McIlroy WE. Control of rapid limb movements for balance recovery: age-related changes and implications for fall prevention. *Age Ageing*. 2006;35:ii12-ii18. <https://doi.org/10.1093/ageing/af1078>
- Maki BE, McIlroy WE. The role of limb movements in maintaining upright stance: the "Change-in-Support" strategy. *Phys Ther*. 1997;77:488-507. <https://doi.org/10.1093/ptj/77.5.488>
- Maki BE, McIlroy WE, Perry SD. Influence of lateral destabilization on compensatory stepping responses. *J Biomech*. 1996;29:343-353. [https://doi.org/10.1016/0021-9290\(95\)00053-4](https://doi.org/10.1016/0021-9290(95)00053-4)
- Scott E, Eagle M, Mayhew A, et al. Development of a functional assessment scale for ambulatory boys with Duchenne muscular dystrophy. *Physiother Res Int*. 2012;17:101-109. <https://doi.org/10.1002/pri.520>
- Pane M, Coratti G, Brogna C, et al. Upper limb function in Duchenne muscular dystrophy: 24 month longitudinal data. *PLoS One*. 2018;13:e0199223. <https://doi.org/10.1371/journal.pone.0199223>
- Muntoni F, Domingos J, Manzur AY, et al. Categorising trajectories and individual item changes of the north star ambulatory assessment in patients with Duchenne muscular dystrophy. *PLoS One*. 2019;14:e0221097. <https://doi.org/10.1371/journal.pone.0221097>
- White MK, Leffler M, Rychlec K, et al. Adapting traditional content validation methods to fit purpose: an example with a novel video assessment and training materials in Duchenne muscular dystrophy (DMD). *Qual Life Res*. 2019;28:2979-2988. <https://doi.org/10.1007/s1136-019-02245-2>
- Leffler MG, Contesse MG, Staunton H. Home-based video assessment of the quality of movement of patients with Duchenne: interviews with physical therapists to inform task selection [abstract]. In: MDA Clinical and Scientific Conference; 2019 Apr 13-17; Orlando, FL, Abstract P.161.
- Contesse MG, Lowes LP, White MK, Dalle Pазze L, McSherry C, Leffler MG. 10th European conference on rare diseases & orphan products (ECRD 2020). *Orphanet J Rare Dis*. 2020;15:310, P14. <https://doi.org/10.1186/s13023-020-01550-1>
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
- Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Chan YH. Biostatistics 104: correlational analysis. *Singapore Med J*. 2003;44:614-619.
- Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med*. 2018;18:91-93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Steffensen BF, Hyde SA, Attermann J, Mattsson E. Reliability of the EK scale, a functional test for non-ambulatory persons with Duchenne dystrophy. *Adv Physiother*. 2002;4:37-47. <https://doi.org/10.1080/140381902317303195>
- Florence JM, Pandya S, King WM, et al. Clinical trials in Duchenne dystrophy. Standardization and reliability of evaluation procedures. *Phys Ther*. 1984;64:41-45. <https://doi.org/10.1093/ptj/64.1.41>
- Pane M, Mazzone ES, Fanelli L, et al. Reliability of the performance of upper limb assessment in Duchenne muscular dystrophy. *Neuromuscul Disord*. 2014;24:201-206. <https://doi.org/10.1016/j.nmd.2013.11.014>
- McDonald CM, McDonald DA, Bagley AM, et al. Relationship between clinical outcome measures and parent proxy reports of health-related quality of life in ambulatory children with Duchenne muscular dystrophy. *J Child Neurol*. 2010;25:1130-1144. <https://doi.org/10.1177/0883073810371509>
- Henricson E, McDonald C. Five-year longitudinal UC Davis CINRG Duchenne natural history study (DNHS) data show mobility-focused POSNA PODCI items are sensitive to 12-month disease progression across all stages of DMD functional ability. *Neuromuscul Disord*. 2015;25:S303. <https://doi.org/10.1016/j.nmd.2015.06.416>
- U.S. Food & Drug Administration. *Duchenne Muscular Dystrophy and Related Dystrophinopathies: Developing Drugs for Treatment*. Silver Spring, MD; FDA; 2018.

26. European Medicines Agency. *Guideline on the Clinical Investigation of Medicinal Products for the Treatment of Duchenne and Becker Muscular Dystrophy*. London, UK; EMA; 2015.
27. Szegarty CA-K, Spitali P. Biomarkers of Duchenne muscular dystrophy: current findings. *Degener Neurol Neuromuscul Dis*. 2018;8:1-13. <https://doi.org/10.2147/DNND.S121099>
28. Aartsma-Rus A, Ferlini A, Vroom E. Biomarkers and surrogate endpoints in Duchenne: meeting report. *Neuromuscul Disord*. 2014;24:743-745. <https://doi.org/10.1016/j.nmd.2014.03.006>
29. U.S. Food & Drug Administration. *Surrogate Endpoint Resources for Drug and Biologic Development*. Silver Springs, MD; FDA; 2018.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Contesse MG, Sapp ATL, Apkon SD, Lowes LP, Dalle Pазze L, Leffler MG. Reliability and construct validity of the Duchenne Video Assessment. *Muscle & Nerve*. 2021;64(2):180-189. <https://doi.org/10.1002/mus.27335>