

Moving towards meaningful measurement: Rasch analysis of the North Star Ambulatory Assessment in Duchenne muscular dystrophy

ANNA MAYHEW¹ | STEFAN CANO² | ELAINE SCOTT³ | MICHELLE EAGLE¹ | KATE BUSHBY¹ |
FRANCESCO MUNTONI⁴ | ON BEHALF OF THE NORTH STAR CLINICAL NETWORK FOR PAEDIATRIC
NEUROMUSCULAR DISEASE

1 Institute of Human Genetics, International Centre for Life, Newcastle University, Newcastle upon Tyne. **2** Peninsula College of Medicine and Dentistry, Plymouth. **3** Muscular Dystrophy Campaign, London. **4** Dubowitz Neuromuscular Centre, Institute of Child Health, University College London, London, UK.

Correspondence to Dr Anna Mayhew at Institute Human Genetics, International Centre for Life, Newcastle University, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK.
E-mail: anna.mayhew@ncl.ac.uk

PUBLICATION DATA

Accepted for publication 19th January 2011.
Published online 17th March 2011.

ABBREVIATIONS

DMD	Duchenne muscular dystrophy
NSAA	North Star Ambulatory Assessment
PSI	Person Separation Index
RUMM 2020	Rasch Unidimensional Measurement Model

AIM Reliable measurement of disease progression and the effect of therapeutic interventions in Duchenne muscular dystrophy (DMD) require clinically meaningful and scientifically sound rating scales. Therefore, we need robust evidence to support such tools. The North Star Ambulatory Assessment (NSAA) is a promising, clinician-rated scale with potential uses spanning clinical practice and clinical trials. In this study, we used Rasch analysis to test its suitability in these roles as a measurement instrument.

METHOD NSAA data from 191 ambulant boys (mean age at assessment 7y 8mo, SD 2y 4mo; range 3y 6mo–15y 5mo) with a confirmed diagnosis of DMD were examined for psychometric properties including clinical meaning, targeting, response categories, model fit, reliability, dependency, stability, and raw to interval-level measurement. All analyses were performed using the Rasch Unidimensional Measurement Model.

RESULTS Overall, Rasch analysis supported the NSAA as being a reliable (high Person Separation Index of 0.91) and valid (good targeting, little misfit, no reversed thresholds) measure of ambulatory function in DMD. One item displayed misfit (lifts head, fit residual 6.9) and there was evidence for some local dependency (stand on right/left leg, climb and descend box step right/left leg, and hop on right/left leg, residual correlations >0.40), which we provide potential solutions for in future use of the NSAA. Importantly, our findings supported good clinical validity in that the hierarchy of items within the scale produced by the analyses was supported by clinical opinion, thus increasing the clinical interpretability of scale scores.

INTERPRETATION In general, Rasch analysis supported the NSAA as a psychometrically robust scale for use in DMD clinical research and trials. This study also demonstrates how Rasch analysis is a useful instrument to detect and understand the key measurement issues of rating scales.

Duchenne muscular dystrophy (DMD) is a severe, genetic X-linked disease that affects 1 in 3600 to 6000 live male births.^{1–3} The predominant feature of the disease is progressive muscle weakness that in the early years manifests itself as delayed motor milestones and inability to run and jump properly. The weakness progresses so that in untreated individuals the ability to walk is lost, on average, at the age of 9 years, but this can be delayed with corticosteroid therapy.^{4,5}

The need to measure physical ability accurately in DMD has become increasingly important, especially as therapies such as daily or intermittent steroids are now an accepted part of standard care⁴ and innovative drugs are trialled. This has made rating scales central to clinical decision making for this group of patients. The North Star Ambulatory Assessment (NSAA) is a widely used 17-item rating scale specifically developed by the North Star Clinical Network for Paediatric

Neuromuscular Disease to measure function in ambulant children with DMD.⁶ The scale has been in use nationally in the UK since 2005 and is currently also used in several other countries and international clinical trials.^{7,8} Thus, data from the NSAA are becoming integral to patient care, prescribing, and policy making. It is, therefore, essential that this scale provides scientifically robust, clinically meaningful, and clinically interpretable results.

The increasing importance of rating scales such as the NSAA highlights the central role of psychometric methods in scale development and testing. In brief, psychometric methods aim to ascertain whether it is legitimate to produce total scale scores from items and the extent to which these scale scores are free from random error (reliability) and measure the attributes they purport to measure (validity). There are two main types of psychometric method: ‘traditional’ and ‘modern’.⁹

Traditional methods are the most commonly used analyses for examining scale reliability and validity.^{10,11} However, although traditional psychometric methods are widely used, they have important limitations for evaluating the measurement properties of rating scales.⁹ These include the following: (1) the data they generate are ordinal rather than interval; (2) scores for people and samples are scale dependent; and (3) scale properties such as reliability and validity are sample dependent. These important limitations can be addressed by applying modern psychometric methods, such as Rasch analysis.¹²

Fundamentally, 'modern' methods focus on the relation between a person's measurement and their probability of responding to an item, rather than the relation between a person's measurement and their observed scale total score. Among the many benefits proffered by this approach is that it leads to the legitimate summing of items to produce total scores and, in turn, total scores produce interval-level measures from ordinal-level rating scale data.^{13,14} This can help to improve the accuracy with which clinical change can be measured. In addition, these methods provide estimates for patients (and items) that are independent of the sampling distribution of items (and patients). This allows for accurate estimates suitable for individual person measurement.

So far, although the NSAA has undergone some limited traditional reliability (interrater, test-retest reliability) and validity analyses,¹⁵ it has yet to undergo a comprehensive psychometric analysis. Given the clinical benefits of new psychometric methods, in this study we aimed to analyse the NSAA using Rasch measurement methods.

METHOD

Data collection

Eligible participants (ambulant boys with a confirmed diagnosis of DMD) were recruited through the North Star Database, a secure web-based database which collects defined medical and physiotherapy assessment data on ambulant boys with DMD from a network of 17 specialist centres in the UK. A request to extract the data from the system was approved by the North Star Governance Committee.

North Star Ambulatory Assessment measure

The NSAA is a multiple item rating scale (17 items) with three ordered response categories (2, 1, or 0) which are summed to give a total score. Items are scored either 2 ('normal' with no obvious modification of activity), 1 (modified method but achieves goal independent of physical assistance from another), or 0 (unable to achieve independently). A total 'ambulatory function' score is generated by summing items. A higher score indicates better motor function. Full test details are available at http://www.muscular-dystrophy.org/how_we_help_you/for_professionals/clinical_databases.

Psychometric evaluation using Rasch analysis

Essentially, a Rasch analysis examines the extent to which the observed data (in this instance physiotherapists' ratings on scale items) 'fit' with predictions of those ratings from the Rasch model (which defines how a set of items should perform

What this paper adds

- Rasch analysis supports the NSAA as a psychometrically sound measure of ambulation in DMD.
- Our findings support good clinical validity, which will help clinicians interpret NSAA scores.
- This paper illustrates the added value of using Rasch analysis for examining rating scales.

to generate reliable and valid measurements).¹⁶ Thus, the difference between expected and observed scores indicates the degree to which valid measurement is achieved. Further detail about Rasch measurement methods are provided elsewhere.^{12,16} We examined eight tests for reliable and valid measurement, which were as follows.

Clinical meaning

We examined the clinical validity of all items within the NSAA to judge the extent to which they were clinically cohesive (contributed to the construct they set out to measure) and ordered in terms of difficulty. We compared this to the Rasch output, which also ranks the items. Five expert neuromuscular physiotherapists, regularly using the NSAA, ranked the 17 items in order of difficulty to establish task hierarchy from a clinical point of view. This enabled a comparison to be made between the Rasch item hierarchy and clinical expectation with the view to check consistency. This was examined qualitatively and statistically (Spearman's *rho*).¹⁷

Targeting

Scale-to-sample targeting concerns the match between the range of ambulatory function measured by the NSAA and the range of ambulation measured in the sample of children.^{18,19} In brief, this was achieved by an examination of the spread of person and item locations in these two relative distributions. This analysis informs us as to how suitable the sample is for evaluating the NSAA and how suitable it is for measuring the sample. Better targeting equates to a better ability to interpret the psychometric data with confidence.

Response categories

Each NSAA item has multiple response categories which reflect an ordered continuum of better ambulatory function (2, 1, and 0). Although this ordering may appear clinically sensible at the item level, it must also work when the items are combined to form a set. Rasch analysis tests this statistically and graphically by threshold locations and plots.²⁰ In brief, when the response options are working as expected, this provides some important evidence for the validity of the scale.

Fit

The items of the NSAA must work together (fit) as a conformable set both clinically and statistically. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of ambulatory motor function. When items do not work together (misfit) in this way, the validity of a scale is questioned. We examined four indicators: (1) log residuals (item-person interaction), (2) χ^2 values (item-trait interaction), (3) *t*-test for unidimensionality,

and (4) item characteristic curves.¹⁸ As there are no absolute criteria for interpreting fit statistics, it is more meaningful to interpret them together, and in the context of their clinical usefulness as an item set.

Person Separation Index

The Person Separation Index (PSI)²¹ is a reliability statistic, comparable to Cronbach's alpha.²² It quantifies the error associated with the measurements of participants in this sample. Higher values indicate greater reliability.

Dependency

The responses to one NSAA item should not directly influence the response to another.^{18,19} If this happens, measurement estimates can be biased and reliability (PSI) is artificially elevated. Rasch analysis determines this effect by examining the residual correlations.

Stability (differential item functioning)

It is important that the NSAA items perform in a similar way across subgroups of children that we would want to compare (e.g. different treatment regimes). The degree to which item performance remains stable across subgroups is known as differential item functioning. In our analysis, we tested treatment type comparing steroid regimes (daily, intermittent regime, no steroids).

Raw score to interval-level measurement

It is important to understand the implication of using raw summed NSAA scores, which are by definition ordinal- as opposed to interval-level measurement. The better the NSAA raw scores approximate interval-level measurement, the more confident we can be in treating one as the other. Rasch analysis estimates linear measurements from raw scores in a graphical plot.

All Rasch analyses were performed using Rasch Unidimensional Measurement Model (RUMM2020) software.²³ A partial credit was used for these data.

RESULTS

Sample

Consent was given before data were entered onto the database. All 191 initial records were eligible for inclusion. These were defined as ambulant boys attending one of the 17 specialist centres in the UK. Of these 191 individuals, 187 complete records from the boys' first assessments were entered onto RUMM2020. The mean age of the sample at assessment was 7 years 8 months (SD 2y 4mo, range 3y 6mo–15y 5mo). For steroid regime at the time of assessment, 77 boys were on a daily regime, 85 were on some form of intermittent regime (mostly 10 days on, 10 days off), and 25 were not on steroids. For this study, the type of steroid was not defined; almost all boys, however, were treated with prednisolone, only a few with deflazacort.

Rasch analysis

Overall, the NSAA performed well against the battery of psychometric tests. The key findings from these tests for reliable and valid measurement were as follows.

Clinical meaning

The ranking of the items by expert opinion compared with the ordering of items using Rasch analysis was similar ($\rho=0.80$); items such as stand, walk, gets to sitting, and standing on one leg are defined as easier by both Rasch analysis and the experts, and run, jump, and hop are considered more difficult. This supports the clinical validity and, therefore, interpretability of NSAA total scores.

Targeting

The item locations spread out (-2.43 to $+2.54$), indicating that the NSAA defines a good continuum with little overlap (i.e. very few items measure the same level of ability; Table I). However, the range of the upper histogram horizontally that extends outside the range of the lower histogram suggests minimal floor and ceiling effect of our current items (Fig. 1).

Table I: Individual item fit for the 17 items in threshold location: order of difficulty, easiest to most difficult

Item	Item number	Item location	Standard error	Fit residual	χ^2	χ^2 p values
Stand	I0001	-2.429	0.189	0.808	4.066	0.131
Walk	I0002	-1.836	0.169	1.198	0.852	0.653
Rise from chair	I0003	-1.252	0.163	2.374	0.132	0.936
Gets to sitting	I0010	-1.15	0.166	1.982	3.168	0.205
Stand on one leg left	I0005	-0.745	0.153	-0.024	2.422	0.298
Stand on one leg right	I0004	-0.609	0.152	0.615	3.257	0.196
Climb box step right	I0006	-0.59	0.153	-3.231 ^a	10.111	0.006
Climb box step left	I0007	-0.224	0.143	-2.256	4.452	0.108
Descend box step right	I0008	-0.163	0.143	-2.626 ^a	11.41	0.003
Descend box step left	I0009	-0.163	0.143	-1.609	2.326	0.312
Lifts head	I0012	0.216	0.13	6.984 ^a	36.578	0
Jump	I0014	0.551	0.123	-1.707	4.723	0.094
Run	I0017	1.01	0.14	-1.545	5.249	0.072
Rise from floor	I0011	1.13	0.18	-0.718	3.488	0.175
Stand on heels	I0013	1.585	0.126	0.389	6.183	0.045
Hop left leg	I0016	2.218	0.134	-0.396	5.402	0.067
Hop right leg	I0015	2.454	0.14	-1.226	3.631	0.163

^aItems that 'misfit' the overall scale (items should lie in range of SD 2.5). Only 'Lifts head' significantly misfits.

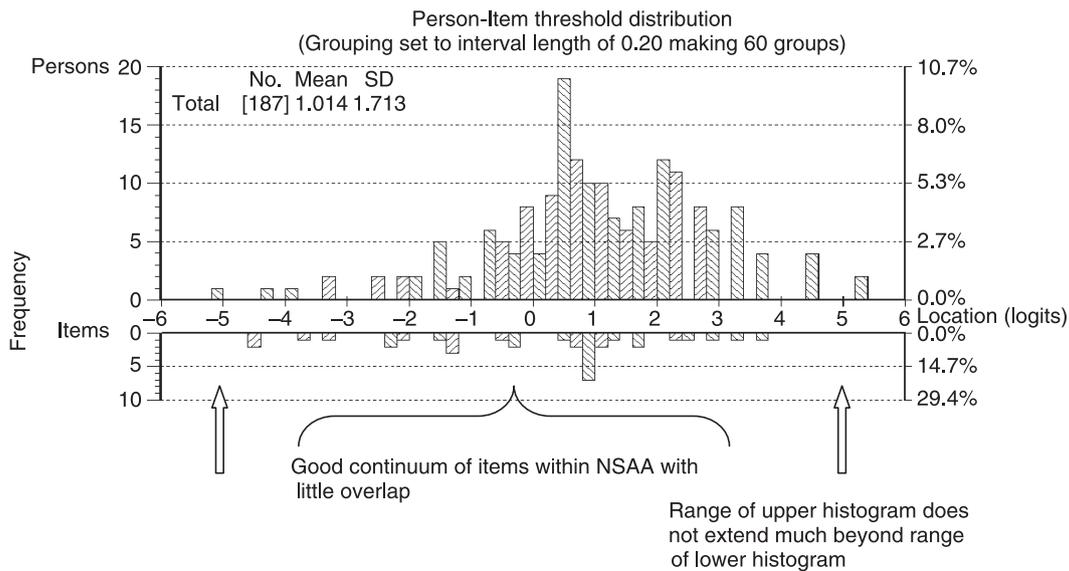


Figure 1: Person-item location distribution. Targeting of the patient sample (top) to the items (bottom). The figure shows the adequate targeting between the distribution of person measurements (upper histogram) and the distribution of item locations (lower histogram). The ceiling/floor effects are minimal as the range of the person measurements (upper histogram 'blocks') closely matched the item locations (lower histogram 'blocks').

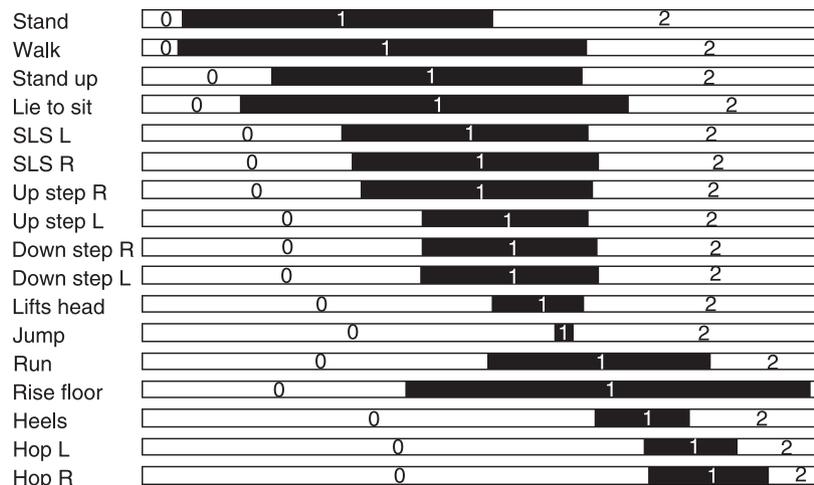


Figure 2: Threshold map for items in ranked order of difficulty according to Rasch analysis. 0, response category labelled 0; 1, response category labelled 1 (black block); 2, response category labelled 2. It would be expected that as a boy's ability increases, he would be more likely to obtain a higher score and that this would increase systematically in a logical progression so that as ability increases he is more likely to score a 0, then a 1, then a 2. This was the case for all the items within the North Star Ambulatory Assessment (i.e. there were no reversed thresholds). SLS, single leg stance; R, right; L, left.

Response categories

The item response option thresholds were ordered for 17 out of 17 items, indicating that the proposed scoring function was working as intended for all items (Fig. 2). This is supported by examination of the category probability curves, which are also illustrated (Fig. 3).

Fit

The overall item-trait interaction χ^2 value was 107.45 (34 df). Three out of 17 items had fit residuals outside the recommended range (-2.50 to 2.50; Table I). Two items exhibited

slight misfit: climb box step right (-3.23) and descend box step right (-2.63). Only one item had notable misfit (lifts head from supine, 6.98), which also had a significant χ^2 probability ($p=0.001$; Table I). Examinations of the graphical indicator of fit (item characteristic curves) suggested this item underdiscriminated. This implies the boys' response to this item was not consistent with those predicted by the Rasch model, unlike other items such as 'jump', which showed good fit of the observed scores with the expected scores (Fig. 4). Unidimensionality was acceptable (t -test 8.2%, binomial test lower 95% confidence interval proportion, 0.05).

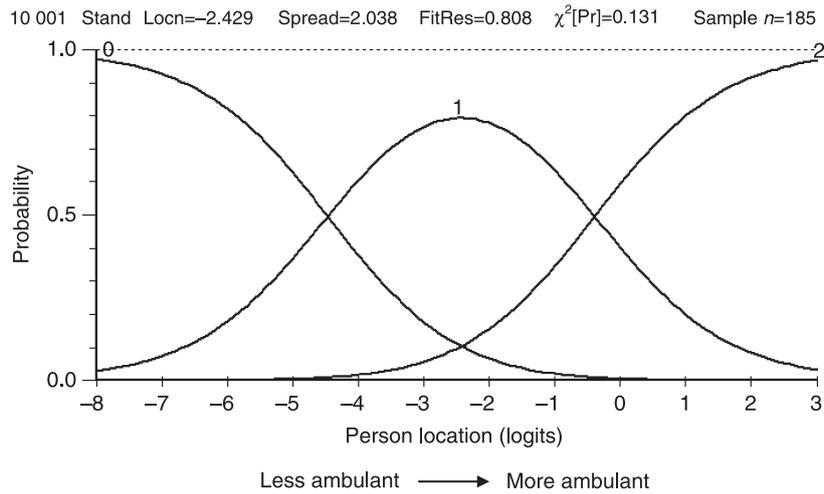


Figure 3: Category probability curve for item 1, 'stand', showing logical progressive order. The x axis symbolizes the construct, with the ambulant ability increasing to the right. The y axis shows the probability of scoring the categories: 0, unable; 1, adapted method but able; 2, able in 'normal' way. Each of the three categories emerged as the most likely to be selected at some point upon the underlying ambulation scale. Locn, location; FitRes, Fit residual; Pr, probability.



Figure 4: Illustrating classic fit of item 14, 'jump'. Curved line, the expected scores for this item; dots, the observed scores for the class intervals at the different level of ability. For item 14, 'jump', this illustrates a 'classic fit' as the dots match the expected scores. Locn, location; FitRes, Fit residual; Pr, probability.

Person Separation Index

The findings from the Rasch analysis indicated that scale reliability was supported by a high Person Separation Index (0.91).

Dependency

Four pairs of items had residuals that were highly correlated (>0.40), implying that a response to one influenced the response to the other: items 3 and 4, 'stand on right leg and left leg'; items 6 and 7 and 8 and 9, 'climb and descend box step right and left'; and items 15 and 16 'hop on right and left leg'. All these pairs appear sequentially in the scale, further supporting this view. This implies that this is a dependency/response bias to an ordering effect.

Stability

Differential item functioning showed that there were no non-uniform differential item functioning issues for steroids. This implies that the item locations were stable regardless of steroid regime. There was a statistically significant uniform differential item functioning issue for just one item: ability to stand ($p < 0.001$; Table II).

Raw score to interval-level measurement

Although the graph appears linear within the central portion, the relation between raw and interval-level measurement is S-shaped and implies a one-point change in NSAA raw scores varies across the scale. This was greatest at the extremes and least at the centre (Fig. 5). For example a change in the raw

Table II: Differential item functioning for steroid regime

Item	Uniform differential item functioning				Non-uniform differential item functioning				
	MS	F	df	p	MS	F	df	p	
I0001	Stand	7.96119	7.45072	2	0.001 ^a	1.06867	1.00014	4	0.409
I0002	Walk	5.05444	4.65374	2	0.011	0.08222	0.0757	4	0.990
I0003	Rise from chair	2.30698	1.75753	2	0.176	1.34797	1.02693	4	0.395
I0004	Stand on one leg, right	1.42023	1.39805	2	0.240	1.09379	1.07671	4	0.370
I0005	Stand on one leg, left	2.89903	3.19346	2	0.043	1.09484	1.20604	4	0.310
I0006	Climb box step, right	2.74303	5.66233	2	0.004	1.48514	3.06571	4	0.018
I0007	Climb box step, left	2.47656	4.27636	2	0.015	1.47214	2.542	4	0.042
I0008	Descend box step, right	0.67096	1.23093	2	0.295	1.3123	2.40753	4	0.051
I0009	Descend box step, left	1.48747	2.08387	2	0.128	1.00361	1.406	4	0.234
I0010	Gets to sitting	0.40816	0.34261	2	0.710	0.94208	0.7908	4	0.533
I0011	Rise from floor	2.31884	2.91005	2	0.057	1.33629	1.67698	4	0.157
I0012	Lifts head	0.32983	0.10712	2	0.898	3.23353	1.05019	4	0.383
I0013	Heels	2.01928	2.03952	2	0.133	0.7679	0.77559	4	0.542
I0014	Jump	1.39328	2.27094	2	0.106	-0.05221	-0.0851	4	0.999
I0015	Hop, right leg	1.62911	2.65686	2	0.073	0.33268	0.54257	4	0.705
I0016	Hop, left leg	3.33942	4.22003	2	0.016	0.48136	0.6083	4	0.657
I0017	Run	0.62887	0.86391	2	0.423	0.98461	1.3526	4	0.253

^aOne significant difference for uniform differential item functioning. MS, mean square; df, degrees of freedom.

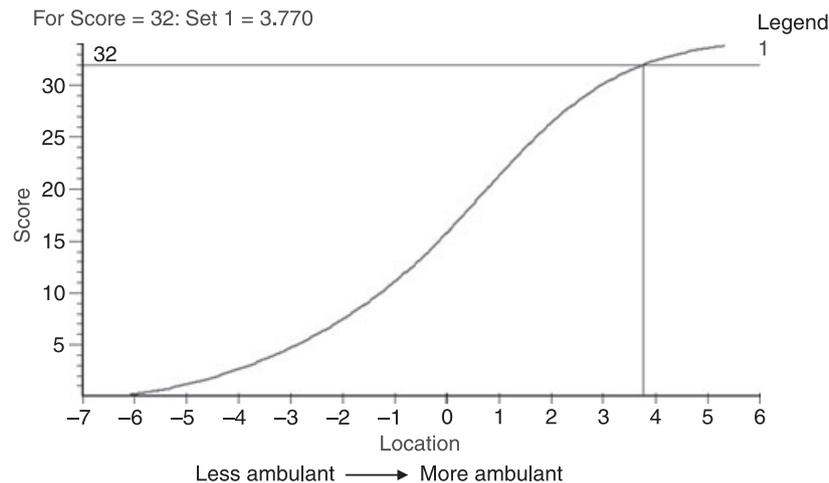


Figure 5: Raw score to interval measure transformation graph. Plot shows the North Star Ambulatory Assessment (NSAA) total scores (y axis) against the intervalized measurements in logits that they imply (x axis). The curve is S-shaped, which means the change in interval measurements associated with a one-point change in the NSAA total score varies across the range of the scale.

score of five points from 0 to 5 is equal to 3.23 logits (logits refers to a person location based on the transformed raw score according to the Rasch model), from 15 to 20 is equal to 0.92 logits, and from 29 to 34 is equal to 2.67 logits.

DISCUSSION

The aim of this study was to evaluate psychometrically the NSAA as an ambulatory outcome measure in a sample of children with DMD, based on Rasch analysis. These preliminary findings imply that, overall, the NSAA was a scientifically robust measure in our sample. Criteria for reliable and valid measurement were, in general, met. In the instances where items were found to be problematical, the issues were minor and could be handled empirically. As such, results supported

the summing of item scores to give a single ambulatory score. From a clinical perspective, this implies that NSAA scores can be interpreted confidently as reliable and valid indicators of ambulatory problems. As the sophisticated techniques offered by Rasch analysis were used, we were able to gain a deep level of understanding of how the NSAA scale, items, and response categories perform. This provides a springboard to propose its use in future clinical research.

Clinical interpretability of the measurements produced by the NSAA was supported by the hierarchical item ordering, which reflected clinical expectation. Items such as stand, walk, and rise from chair were considered easier, and items such as a jump, hop, and run were considered more difficult. This was found to be consistent with the Rasch output. Other items that

therapists found more difficult to rate in terms of a patient's performance were stand on heels and lifts head. The ability to stand on heels is especially dependent upon the available range of ankle dorsiflexion, which is often compromised in DMD owing to contractures in the tendo-achilles. The severity of contractures and response to therapeutic strategies may follow a variable pattern within this group of patients, thus making this item problematical to predict in the hierarchy of 'difficulty' in the NSAA.^{24,25} The ability to lift one's head from supine may remain, despite deterioration in other functional abilities with age, or it may never be achievable despite good lower limb function. This makes clinical interpretation of its difficulty an issue, and supports the finding that this item failed so many of the psychometric tests, in particular demonstrating notable misfit. Although both these items are included in the scale for clinical purposes, it may be that further consideration needs to be given to their exclusion for clinical trials.

Scale-to-sample targeting was good and there were minimal floor and ceiling effects (only two out of 187 participants scored a maximum of 34). It may be appropriate to increase the measurement range of the NSAA and consider the inclusion of items that relate to a higher functional ability such as repeated hops or a squat (without risking the unidimensionality of the scale). However, within the NSAA there are timed tests of rise from the floor and traverse 10m, which may offer further discrimination of those boys who are achieving higher scores. Given the later age at which boys with DMD who are treated with steroids lose ambulation, rather than add extra items relating to gross motor function, it is perhaps more clinically appropriate to shift the focus from gross motor abilities to the evaluation of activities related to daily living (e.g. the Egen Klassifikation Scale^{26,27}), upper limb function, respiratory, cardiac, or scoliosis monitoring using methods validated for DMD.

The ordering of the response category thresholds in the NSAA showed the scoring worked as intended. Therefore, a more able boy would be more likely to score 2, than 1 or 0, and a less able boy would be more likely to score 1 or 0, than a 2 in each item. This supports the number and type of response options used in the NSAA. It also lends further validity to the fact that the scale is measuring one construct. However, as discussed, there was one significantly misfitting item: lifts head.

The dependency revealed by the high residual correlations between items that compared right and left will require some consideration for future development of the NSAA and could artificially inflate reliability (PSD). However, this dependency between items relating to symmetry would be expected clinically as DMD is not a condition that usually presents with a high degree of asymmetry. Physiotherapy experts recognize the need to evaluate such activities bilaterally so that variations can be addressed therapeutically to prevent potential problems such as worsening asymmetry of contractures. Thus, there is a balance that needs to be made between clinical assessment and measurement requirements. From a measurement point of view, one item from each pair could be included in scoring. However, in clinical practice and assessment, the NSAA can

still include all the items to be used qualitatively. Examinations in other datasets will help clarify this situation.

One key finding is that the NSAA appeared stable across treatment/no treatment subgroups. This suggests that a boy's current steroid regime (not on steroids, an intermittent or continuous regime) does not generally influence the manner in which he scores items. The uniform differential item functioning issue for standing ability could be because steroids are keeping boys on their feet for longer, or it may indicate a confounding variable such as contractures within the tendo-achilles, as already discussed. Given the existence of differential item functioning, there are two options: the item could be deleted or 'split' (i.e. two scores derived from this item to account for the different scoring patterns). The implications of this require further investigation and could in fact be caused by skewed analysis due to the imbalance of the group sizes.

Our findings reveal the need for caution when using NSAA ordinal total raw scores for measuring clinical change. The relation between raw and interval scores is S-shaped, which means that the change in interval measurements associated with a one-point change in the NSAA total score varies up to fourfold across the range of the scale. This emphasizes one of the important reasons of moving towards Rasch-transformed measurements to give us a more precise scale for use in treatment trials, where a one-point change at any point in the scale has an equal value.

Our study has two main limitations. First, the dataset was formed from cross-sectional data. Cross-sectional data were included that provided a sufficient sample for analysis. However, inclusion of longitudinal data would enable stability of the scale over time to be assessed. This is an important issue, considering the length of time over which change occurs in this group. A more in-depth examination of the influence of age would also be appropriate, considering the changing clinical course of the disease with the advent of steroids. Second, validity testing was also limited. In particular, we were restricted in the extent to which we could examine aspects of construct validity. Further examinations would, therefore, be beneficial.

Finally, our study illustrates the value of Rasch analysis, which adds sophistication and refinement to traditional psychometric methods, and provides detailed diagnostic item-level data. In the current context, this added value is clearly highlighted by the detailed information it provides to the reliability and validity of the NSAA. In particular, these methods provide a solid platform upon which to bring together clinical sensibility with statistical analyses. The implication is that this rigorous methodology has shown the NSAA to be reliable and valid, especially demonstrated by the clinical and statistical relevance of the individual scores for all 17 items. These analyses provide an initial evidence base, which creates the potential to transform ordinal NSAA scores into interval-level scores, in turn vastly improving the interpretability of change scores across breadth of the scale. This would enable future researchers to interpret NSAA data and build upon them, if they want to use this tool in future DMD research and clinical trials.

ACKNOWLEDGEMENTS

Collaborators in the North Star Clinical Network for Paediatric Neuromuscular Disease were as follows: AY Manzur (clinical lead); F Muntoni, S Robb, M Main, J Kemp, V Ricotti (Great Ormond Street Hospital, London, UK); E Scott (Muscular Dystrophy Campaign, London, UK); K Bushby, V Straub, A Sarkozy, E Strehle, R Venkateswaran, E Eagle, A Mayhew (Institute of Human Genetics, Newcastle, UK); H Roper, H McMurchie, A Grace (Heartlands Hospital, Birmingham, UK); S Spinty, G Peachey, S Shillington (Alder Hey Children's Hospital, Liverpool, UK); R Quinlivan, L Groves (Robert Jones and Agnes Hunt Royal Orthopaedic Hospital, Oswestry, UK); E Wraige, H Jungbluth, J Sheehan, R Spahr (Evelina Children's Hospital, London, UK); I Hughes, E Bateman, C Cammiss (Royal Manchester Children's Hospital, UK); AM Childs, L Pallant, K Psyden (Leeds General Infirmary, UK); P Baxter (Shef-

field Children's Hospital, UK); K Naismith, A Keddie (Kings Cross Hospital, Dundee, UK); I Horrocks, R McWilliam, M Di Marco (Yorkhill Children's Hospital, Glasgow, UK); L Hartley, B Sheen, J Fenton-May (University Hospital Wales, Cardiff, UK); P Jardine, A Majumdar, L Jenkins (Frenchay Hospital, Bristol, UK); G Chow, A Miah (Queens Medical Centre University Hospital, Nottingham, UK); C de Goede (Preston Royal Hospital, UK); N Thomas, M Geary, K Keslake (Southampton General Hospital, UK); C White, K Greenfield (Morrison Hospital, Swansea, UK); S MacAuley (Royal Belfast Hospital for Sick Children, UK); A Baxter, Y Yirrell, C Longman (Royal Hospital for Sick Children, Western General Hospital, Edinburgh, UK).

Muscular Dystrophy Campaign support and fund the North Star Project. The Medical Research Council Neuromuscular Centre support to the North Star Database is also gratefully acknowledged.

REFERENCES

1. Drousiotou A, Ioannou P, Georgiou T, et al. Neonatal screening for Duchenne muscular dystrophy: a novel semiquantitative application of the bioluminescence test for creatine kinase in a pilot national program in Cyprus. *Genet Test* 1998; **2**: 55–60.
2. Bradley D, Parsons E. Newborn screening for Duchenne muscular dystrophy. *Semin Neonatal* 1998; **3**: 27–34.
3. Emery AE. Population frequencies of inherited neuromuscular diseases – a world survey. *Neuromuscul Disord* 1991; **1**: 19–29.
4. Bushby K, Finkel R, Birnkrant DJ, et al. Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, pharmacological and psychosocial management. *Lancet Neurol* 2010; **9**: 77–93.
5. Manzur AY, Kuntzer T, Pike M, Swan AV. Glucocorticoid corticosteroids for Duchenne muscular dystrophy. *Cochrane Database Syst Rev* 2008; **1**: CD003725.
6. Scott E, Eagle M, Main M, Sheehan J. The North Star Ambulatory Assessment. Poster presented at the Annual Meeting of the British Paediatric Neurology Association, 2006. *Dev Med Child Neurol* 2006; **48**: (Suppl. 104) 27.
7. Kinali M, Arechavala-Gomez V, Feng L, et al. Local restoration of dystrophin expression in Duchenne muscular dystrophy: a single blind, placebo-controlled dose escalation study using morpholino antisense oligomer AVI-4658. *Lancet Neurol* 2009; **8**: 918–28.
8. Shrewsbury SB, Cirak S, Guglieri M, Bushby K, Muntoni F. Current progress and preliminary results with the systematic administration trial of AVI-4658, a novel phosphorodiamidate morpholino oligomer (PMO) skipping dystrophin exon 51 in Duchenne muscular dystrophy (DMD) World Muscle Society Meeting, Kumamoto, Japan 12–16th October, 2010. *Neuromuscul Disord* 2010; **20**: 639–40.
9. Hobart JC, Cano SJ, Zajick JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol* 2007; **6**: 1094–105.
10. Cano SJ, Hobart JC. Watch out, watch out, the FDA are about. *Dev Med Child Neurol* 2008; **50**: 408–9.
11. Novick MR. The axioms and principal results of classical test theory. *J Math Psychol* 1966; **3**: 1–18.
12. Hobart JC, Cano SJ. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Monogr UK Health Technol Assess Prog* 2009; **13**: 1–200.
13. Wright BD, Linacre JM. Observations are always ordinal: measurements, however must be interval. *Arch Phys Med Rehabil* 1989; **70**: 857–60.
14. Wright BD, Stone MH. Best Test Design: Rasch Measurement. Chicago, IL: MESA, 1979.
15. Mazzone ES, Messina S, Vasco G, et al. Reliability of the North Star Ambulatory Assessment in a multicentric setting. *Neuromuscul Disord* 2009; **19**: 458–61.
16. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Education Research, 1960.
17. Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904; **15**: 72–101.
18. Wright BD, Masters G. Rating Scale Analysis: Rasch Measurement. Chicago, IL: MESA, 1982.
19. Hobart JC, Riazi A, Thompson AJ, et al. Getting the measure of spasticity in multiple sclerosis: the Multiple Sclerosis Spasticity Scale (MSSS-88). *Brain* 2006; **129**: 224–34.
20. Hagquist C, Andrich D. Is the Sense of Coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. *Pers Individ Diff* 2004; **36**: 955–68.
21. Andrich D. Distinctions between assumptions and requirements in measurement in the social sciences. In: Keats JA, Taft R, Heath RA, Lovibond SH, editors. Proceedings of the XXIVth International Congress of Psychology. North Holland: Elsevier Science Publications BV, 1989: 7–16.
22. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; **16**: 297–334.
23. RUMM Laboratory. RUMM2020, Version 4.0 for Windows (Upgrade 4600.0109). Perth, Western Australia: RUMM Laboratory, 1997–2007.
24. Scott OM, Hyde SA, Goddard C, Dubowitz V. Prevention of deformity in Duchenne muscular dystrophy. A prospective study of passive stretching and splinting. *Physiotherapy* 1981; **67**: 177–80.
25. Hyde SA, Fløytrup I, Glent S, et al. A randomized comparative study of two methods for controlling tendo achilles contracture in Duchenne muscular dystrophy. *Neuromuscul Disord* 2000; **10**: 257–63.
26. Steffensen B, Hyde S, Lyager S, et al. Validity of the EK scale: a functional assessment of non-ambulatory individuals with Duchenne muscular dystrophy or spinal muscular atrophy. *Physiother Res Int* 2001; **6**: 119–34.
27. Steffensen BF, Lyager S, Werge B, Rahbek J, Mattson E. Physical capacity in non-ambulatory people with Duchenne muscular dystrophy or spinal muscular atrophy: a longitudinal study. *Dev Med Child Neurol* 2002; **44**: 623–32.