RESEARCH PAPER

Measuring arm function early after stroke: is the DASH good enough?

Karen Baker,[1] Louise Barrett,[2] E Diane Playford,[1] Trefor Aspden,[3] Afsane Riazi,[3] Jeremy Hobart[2]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jnnp-2015-310557).

[1]Department of Brain Repair and Rehabilitation, UCL Institute of Neurology, London, UK
[2]Clinical Neurology Research Group, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK
[3]Department of Psychology, Royal Holloway University of London, Egham, Surrey, UK

Correspondence to
Professor Jeremy Hobart, Clinical Neurology Research Group, Plymouth University Peninsula Schools of Medicine and Dentistry, Room N13 ITTC Building, Plymouth Science Park, Derriford, Plymouth PL6 8BX, UK; jeremy.hobart@plymouth.ac.uk

Received 5 February 2015
Accepted 18 June 2015
Published Online First 15 July 2015

To cite: Baker K, Barrett L, Playford ED, *et al. J Neurol Neurosurg Psychiatry* 2016;**87**:604–610.

ABSTRACT

Objective Despite a growing call to use patient-reported outcomes in clinical research, few are available for measuring upper limb function post-stroke. We examined the Disabilities of the Arm, Shoulder and Hand (DASH) to evaluate its measurement performance in acute stroke. In doing so, we compared results from traditional and modern psychometric methods.

Methods 172 people with acute stroke completed the DASH. Those with upper limb impairments completed the DASH again at 6 weeks (n=99). Data (n=271) were analysed using two psychometric paradigms: traditional psychometric (Classical Test Theory, CTT) analyses examined data completeness, scaling assumptions, targeting, reliability and responsiveness; Rasch Measurement Theory (RMT) analyses examined scale-to-sample targeting, scale performance and person measurement.

Results CTT analyses implied the DASH was psychometrically robust in this sample. Data completeness was high, criteria for scaling assumptions were satisfied (item-total correlations 0.55–0.95), targeting was good, internal consistency reliability was high (Cronbach's α=0.99) and responsiveness was clinically moderate (effect size=0.51). However, RMT analyses identified important limitations: scale-to-sample targeting was suboptimal, 4 items had disordered response category thresholds, 16 items exhibited misfit, 3 pairs of items had high residual correlations (>0.60) and 84 person fit residuals exceeded the recommended range.

Conclusions RMT methods identified limitations missed by CTT and indicate areas for improvement of the DASH as an upper limb measure for acute stroke. Findings, similar to those identified in multiple sclerosis, highlight the need for scales to have strong conceptual underpinnings, with their development and modification guided by sophisticated psychometric methods.

INTRODUCTION

Approximately 70–80% of people with acute stroke have upper limb dysfunction.[1 2] This can affect a person's ability to perform activities of daily living (ADLs), including self-care, leisure, work and social activities, and impacts on levels of independence.[3] Therefore, recovery of upper limb function is often a primary goal for rehabilitation following stroke.[4 5] It is commonly assessed using clinician-rated or performance outcome measures (ClinROs, PerFOs), including the Box and Block Test[6] and the Action Research Arm Test.[7]

One goal of patient-reported outcome (PRO) measures is to quantify the activity and participation restrictions that arise from impairments in order to examine their impact on individuals' daily lives and evaluate their treatments.[5 8] However, despite a growing call to include PRO measures in clinical research, including stroke,[9–11] few upper limb PRO measures have been used in acute stroke.[3 5]

The Disabilities of the Arm, Shoulder and Hand (DASH) is a PRO instrument purporting to measure physical function and symptoms in people with upper limb disorders.[12] Originally developed for use in orthopaedic populations,[12] it is one of the most widely used upper limb rating scales.[13 14] Its psychometric properties have been extensively evaluated in musculoskeletal disorders;[3] however, there are no published psychometric evaluations of the DASH in stroke.[3 15]

The widespread use of the DASH underpinned our evaluation in people with acute stroke to examine its suitability as an outcome measure. We used two psychometric paradigms: traditional psychometric methods based on Classical Test Theory (CTT)[16] and modern psychometric methods based on Rasch Measurement Theory (RMT).[17] There were two aims: to evaluate the measurement properties of the DASH in people with acute stroke, and to compare and contrast CTT with RMT.

METHODS

Participants, recruitment and data collection

This is a pooled sample of three subgroups of people in the early post-stroke period. People admitted to the Hyper-acute Stroke Unit (HASU) at the National Hospital for Neurology and Neurosurgery (NHNN) completed the DASH regardless of presence or extent of upper limb dysfunction as part of a routine battery of outcome measures (n=125). Other people with upper limb dysfunction were recruited from the Albany Rehabilitation Unit (ARU) at the NHNN (n=34) and the National Rehabilitation Unit (NRU) at the NHNN (n=13). Participants were between 48 h and 12 weeks post-stroke, 18+ years of age, had an imaging-confirmed diagnosis of stroke, were screened by a research nurse for suitability for inclusion in the study (including cognition and language/communication), and provided full informed consent. People unable to read or with difficulties understanding the questionnaire (due to severe cognitive or language/communication impairment) were excluded. The DASH was administered across all three units by a clinician (the first author) who provided instructions and support during completion. Ethical approval was obtained from the Joint Research Ethics Committee of the Institute of Neurology and the NHNN.

PRO instruments

The DASH has 30 items aiming to measure physical function and symptoms in people with upper limb disorders.[12] Items are scored from 1 (no difficulty) to 5 (unable), summed to generate a total 'disability/symptom' score, and averaged to produce a mean item score between 1 and 5. This value is transformed to a score out of 100 by subtracting one and multiplying by 25. A higher score indicates greater disability.

Analysis plans

We used the approach some of us had taken in examining the DASH in multiple sclerosis (MS).[18] Specifically, we compared and contrasted psychometric evaluations using CTT[16] and RMT.[17]

Traditional psychometric (CTT) methods

CTT methods examined five psychometric properties of the DASH; the statistical methods and associated criteria are documented fully elsewhere.[18–26] Specifically, and in brief, we examined DASH data against published criteria for: (1) data completeness (percent missing data for each item[21]); (2) scaling assumptions (similarity of item means and variances, magnitude and similarity of corrected item-total correlations[19–22]); (3) scale-to-sample targeting (score means, SD, floor and ceiling effects[23 24]); (4) internal consistency reliability (Cronbach's α[25]; mean inter-item correlation[26]); and (5) responsiveness, examined by comparing baseline and 6-week scores (effect size and standardised response mean). Data analyses were conducted using SPSS (V.19.0).

Table 1 Sample characteristics at baseline (n=172)

Variable	Total
Gender, % (n)	
Female	41 (71)
Age, years	
Mean (SD)	61 (17)
Range	18–93
Time post-stroke, weeks	
Mean (SD)	3 (3)
Range	1–12
Handedness, % (n)	
Right	97 (166)
Aphasia, % (n)	
Yes	8 (13)
Upper limb impairment, % (n)	
Yes	58 (99)
No	42 (73)
Treatment group, % (n)	
HASU	73 (125)
ARU	20 (34)
NRU	7 (13)

ARU, Albany Rehabilitation Unit at the National Hospital for Neurology and Neurosurgery (NHNN); HASU, Hyper-acute Stroke Unit at the NHNN; NRU, National Rehabilitation Unit at the NHNN.

RMT methods

In RMT, the degree to which measurements can be derived from item responses is evaluated using a single mathematical equation, the Rasch model, which defines how a set of items should perform in order to generate reliable and valid measurements.[17 27 28] RMT is explained for clinicians elsewhere.[29] Briefly, RMT examines the extent to which the observed scores (person responses) 'fit' the expected values predicted by the Rasch model, indicating the degree to which rigorous measurement is achieved. RMT analyses were grouped into three areas: scale-to-sample targeting, scale performance and person measurement. The methods and associated criteria are documented fully elsewhere.[18 27 29–32] Data analyses were conducted using RUMM2030.[33]

RESULTS

Sample

The sample (n=172) included: 125 people admitted to the HASU, 34 people admitted to the ARU, and 13 people admitted to the NRU (table 1). People with upper limb impairments, as identified from clinical assessment, completed the DASH 6 weeks later (n=99; see online supplementary table 1). Therefore, data from 271 questionnaires were stacked for

Table 2 Measurement characteristics: raw score metric (n=271)

Measurement characteristic*	Value
Data completeness	
Item missing data, %	0
Computable scale scores, %	100
Scaling assumptions	
Item mean scores: range	1.59–3.15
Item SD: range	0.78–1.49
Item variance: range	0.60–2.22
Corrected item-total correlations: range	0.55–0.95
Targeting	
Mean score (SD)	36.0 (26.8)
Possible score range†	0–100
Observed score range	0–88
Ceiling/floor effect, %‡	14.4/0
Skewness	0.03
Reliability	
Cronbach's α	0.99
Mean inter-item correlation	0.70
SEM§	2.68
95% CI¶	±5.25
Responsiveness: Group level comparison (n=99)**	
Time 1 mean (SD) (range)	48.2 (24.3) (0 to 88)
Time 2 mean (SD) (range)	35.8 (23.3) (0 to 85)
Change mean (SD) (range)††	12.4 (17.7) (−41.7 to +68.3)
t Value (p)	7.01 (0.000)
ES‡‡	0.51
SRM§§	0.70

*Measurement characteristics based on raw scores.
†High scores indicate greater disability.
‡Ceiling effect=% scoring 0 (least impact on disability), floor effect=% scoring 100 (greatest impact on disability).
§SEM, SE of measurement=SD$\sqrt{(1-\alpha)}$.
¶95% CI around individual person scores=±1.96×SEM.
**Participants with upper limb impairments only, measured at baseline (Time 1) and 6 weeks (Time 2).
††Change=Time 1−Time 2.
‡‡ES, effect size=mean change/SD Time 1.
§§SRM, standardised response mean=mean change/SD change.

analysis.[29] At baseline, the sample mean age was 61 years (range 18–75), mean time post-stroke was 3 weeks (range 1–12) and 41% were women (table 1).

Traditional psychometric (CTT) methods

The results of CTT analyses supported the DASH as a psychometrically robust measure of upper limb function in stroke. In summary, traditional psychometric criteria were satisfied for all measurement properties evaluated (data completeness, scaling assumptions, targeting, reliability, responsiveness; table 2).

Data completeness

Table 2 shows that data completeness was high: there were no item-level missing data, and scale scores were computable for all respondents.

Scaling assumptions

Criteria for scaling assumptions were satisfied: items had similar mean scores and variances, and all corrected item-total correlations exceeded 0.30 (table 2).

Targeting

DASH scores spanned 88% of the scale range and were not notably skewed. There was no floor effect but a ceiling effect of 14.4%, below the recommended maximum of 20% (table 2).[23 24]

Reliability

Internal consistency reliability was very high (Cronbach's α=0.99), and the mean inter-item correlation (0.70) exceeded the recommended minimum of 0.30 (table 2). The 95% CI around DASH scores was ±5.25 points.

Responsiveness

The mean change in scores for the subsample of people who completed the DASH again at 6 weeks (n=99) was 12.4 points (SD 17.7; table 2). This group-level improvement was statistically significant and clinically moderate according to Cohen's criteria.[34]

RMT methods

RMT analyses were more informative than those from CTT. As some item response categories had not been endorsed, we invoked the null category adjustment feature available in RUMM2030.[35]

Scale-to-sample targeting

Scale-to-sample targeting was suboptimal. Figure 1 shows that the sample appears reasonable for examining scale performance: the sample covers the item locations. However, the scale does not cover the sample: a number of people are not covered by the items and the ceiling effect is notable.

Scale performance

Did the item response categories work as intended?

Thresholds were disordered for 4/30 items implying that the 5-category scoring function was not working as intended for these items (table 4; see online supplementary figure 1). For one item ('write'), people appeared to have difficulty discriminating between the first three categories (see online supplementary figure 1B). For three items ('pain', 'pain performing an activity', 'tingling'), people appeared to have difficulty discriminating between the final three categories (see online supplementary figure 1C).

What continuum was mapped out by the items?

Item locations ranged from −1.61 to +1.78 logits; item thresholds ranged from −3.95 to+3.59 logits, indicating the items mapped out a measurement continuum (table 3). However, there was limited spread and items were bunched at points along the continuum (figure 1).

To what extent did the items work together?

Sixteen items had fit residuals outside the recommended range (−2.5 to+2.5), one notably so ('feeling less capable', +12.39), and eight items had significant χ^2 values (table 4). Examinations of the graphical indictor of fit (ICC) showed that most items displayed reasonable visual fit despite statistical misfit. However, ICCs for three items suggested these were under-discriminating: 'write', 'sexual activities' and 'feeling less capable' (see online supplementary figure 2).

To what extent did the response to one item bias the response to another?

Sixty-five pairs of items had residual correlations exceeding the criteria of <0.30 (15% of total correlations), implying that a response to one item influenced the response to the other item. Three pairs of items correlated highly (>0.60): 'do heavy household chores' with 'garden' (0.63); 'recreational activities: force/impact' with 'recreational activities: move arm freely' (0.80); and 'pain' with 'pain performing an activity' (0.87).

Person measurement

Person measures covered a wide range (−6.65 to+2.90 logits) and the Person Separation Index (PSI=0.96) implied good sample separation and high reliability (table 3). However, 84 person fit residuals (range −5.85 to+5.00) exceeded the recommended range (−2.5 to +2.5), implying that approximately 36% of people gave responses not in keeping with expectations.

Table 3 shows that group-level responsiveness analyses recorded a significant improvement at 6 weeks post-stroke, clinically mild-to-moderate according to Cohen's criteria.[34] At the individual person level, 51% had a statistically significant improvement and a further 30% made a non-significant improvement.

DISCUSSION

Our aim was to evaluate the measurement performance of the DASH in a sample of people early post-stroke. Traditional psychometric methods implied that the DASH performed well as a measure of upper limb function. Our findings were similar to our CTT examination of the DASH in MS[18] and supported the summing of item scores into a single upper limb symptom/disability score. However, RMT analysis provided more sophisticated information and raised concerns about the DASH as an outcome measure in acute stroke.

RMT analyses revealed that, for four items, the scoring function did not work as intended. One explanation is that there are too many response categories for people to reliably choose between. This finding is consistent with our previous study in MS.[18] Another explanation arises from the suboptimal targeting. For the pain items, there were not enough people located at the more disabled end of the continuum to infer confidently that threshold disordering exists. RMT analyses identify problems; they do not indicate the cause. Exploring possible reasons for the disordering is important as ordered thresholds are necessary for scale validity.[30 36]

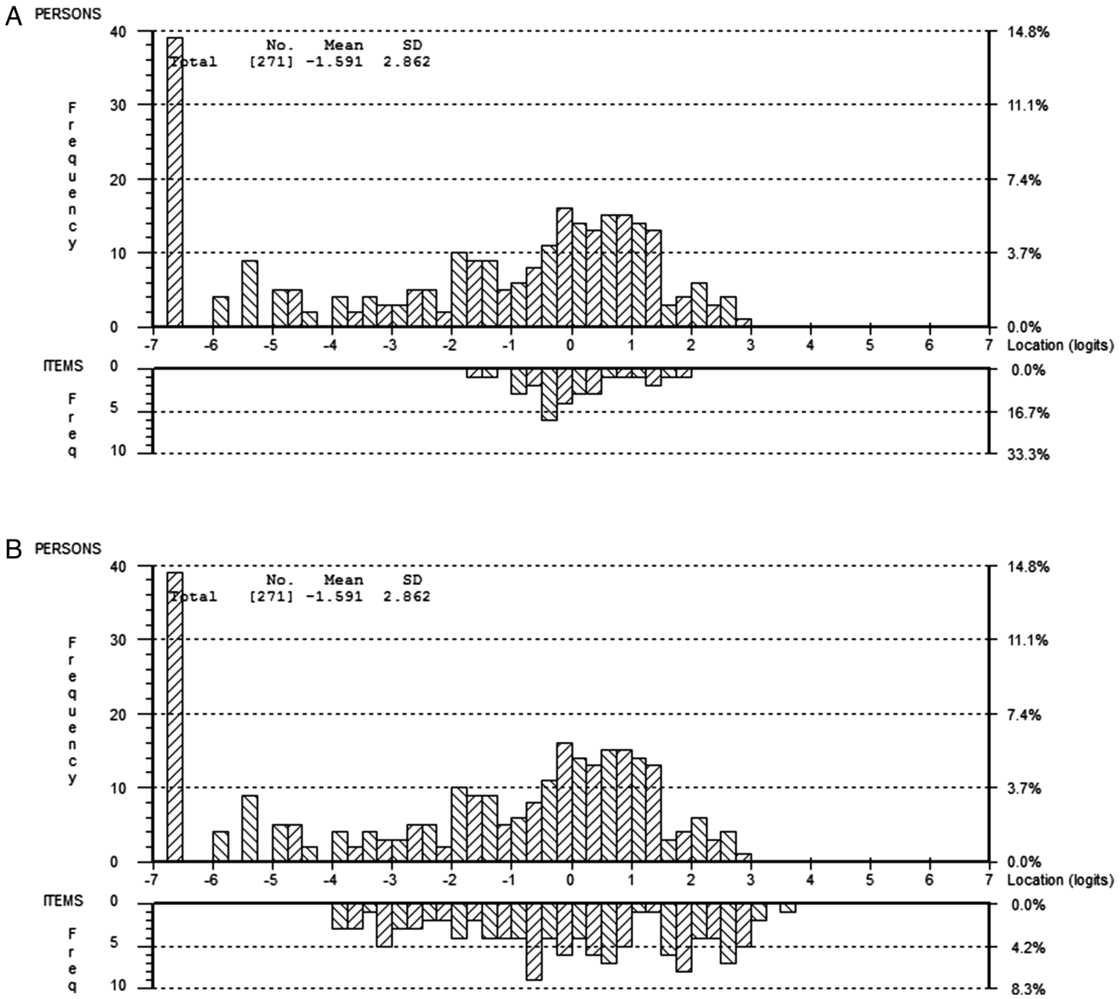

Figure 1 Targeting of the sample to Disabilities of the Arm, Shoulder and Hand (DASH) items. (A) This figure shows the distribution of person measurements (upper histogram) against the distribution of item locations (lower histogram). People (upper histogram) are located along the continuum from more ability (left hand bars) to less ability (right hand bars). Items (lower histogram) are located relative to each other: the easiest items (requiring less ability to perform) represented by the bars on the right and the most difficult items (requiring more ability to perform) represented by the bars on the left. People located outside the scale's measurement range (−2 to+2 logits) indicate suboptimal scale-to-sample targeting. (B) This figure shows the distribution of person measurements (upper histogram) against the distribution of the item *threshold* locations (lower histogram). The lower histogram shows the distribution of item thresholds which represent the boundaries between adjacent response categories. A threshold is the point on the continuum at which a response in either of two adjacent categories is equally likely. The DASH items have five response categories, so there are four boundaries or thresholds for each item. The item location (lower histogram in figure 1A) is the mean of the four threshold locations.

Over 50% of items had statistical misfit implying that the items were not as statistically cohesive as required for deriving measurements from an item set. Item misfit has many possible causes, including disordered thresholds. However, misfit may arise from the content of the item set. The DASH has items measuring physical function, symptoms and social participation. Previously, we suggested the DASH is capturing a broader construct, not simply upper limb functioning.[18] From a measurement perspective, combining symptoms with functioning threatens scale validity. From a clinical perspective, this means that DASH scores lack meaning and interpretability.

The hierarchical ordering of the DASH items suggests, from a clinical perspective, that the scale items represent more than one dimension (multidimensionality). Examination of the item locations in table 4 reveals that the item ordering is not intuitive clinically. We explored this further by combining the items into four clinically sensible groups—symptoms (items 24–29), participation (items 17–23, 30), dexterity (items 1–4, 15, 16) and power/range of motion (items 5–14)—and performing an exploratory subtest analysis. Subtest analyses are performed post hoc and can be used to explore the presence of multidimensionality within a scale. Perhaps counterintuitively, traditional reliability indicators can be over-inflated by multidimensionality, in part because these are not indicators of unidimensionality as is mistakenly thought.[31] In a subtest analysis, items are grouped into 'subtests' which are then treated by the analysis as single 'super' items. So here, the four item groups or 'subtests' are analysed as if they were a four-item scale. If reliability indicators (eg, PSI and α) fall, multidimensionality is implied. We found that the PSI dropped from 0.96 to 0.87 (α dropped from 0.99 to 0.92), implying that DASH reliability was artificially inflated, supporting our clinical impression of multidimensionality. However, this issue is not simple because a fall in reliability indicators following subtest analysis can also occur when there is item response bias (local dependence). Also, α values are dependent on the number of extreme scores (here a notable

Table 3 Measurement characteristics: Rasch measurement metric (n=271)

Measurement characteristic*	Value
Item locations	
Mean (SD)	0.00 (0.85)
Range	−1.61 to +1.78
Thresholds	
Range	−3.95 to +3.59
Person measures	
Mean (SD)	−1.59 (2.86)
Range	−6.65 to +2.90
Reliability	
Person Separation Index†	0.96
Responsiveness: group level comparison (n=99)‡	
Time 1 mean (SD) (range)	−0.36 (2.39) (−6.65 to +2.90)
Time 2 mean (SD) (range)	−1.28 (2.22) (−6.65 to +2.57)
Change mean (SD) (range)§	0.92 (1.85) (−4.66 to +5.49)
t value (p)	4.96 (0.000)
ES¶	0.38
SRM**	0.50
Responsiveness: individual person level comparison (n=99)††	
Significant improvement, % (n)	51 (50)
Non-significant improvement, % (n)	29 (29)
No change, % (n)	7 (7)
Non-significant worsening, % (n)	6 (6)
Significant worsening, % (n)	7 (7)

*Measurement characteristics based on raw score transformation into linear measurements.
†Person Separation Index (PSI), a reliability statistic analogous to Cronbach's α.
‡Participants with upper limb impairments only, measured at baseline (Time 1) and 6 weeks (Time 2).
§Change=Time 1−Time.
¶ES, effect size=mean change/SD Time 1.
**SRM, standardised response mean=mean change/SD change.
††Significant improvement=SigChange≥+1.96; Non-significant improvement=0<SigChange+1.95; No change=SigChange=0; Non-significant worsening=−1.95<SigChange<0; Significant worsening=SigChange≤−1.96
SigChange=change/SED where SED, SE of the difference=$\sqrt{(SE\,T1)^2+(SE\,T2)^2}$.

amount) and item–item correlations (here very high). We would stress that this issue is complex and calls for a full discussion and examination beyond the scope of this manuscript, and therefore, refer interested readers to other sources.[31 32]

As highlighted above, reliability indicators can also be inflated by local dependence among items, or item response bias.[31] We found 65 pairs of items with residual correlations >0.30 and three pairs of items which correlated highly (>0.60) suggesting local dependency. All three pairs of highly correlated items appear sequentially in the DASH and have similar content, implying that this dependency may be due to an ordering and/or content effect.

Targeting is worthy of particular attention when a scale is being used outside its original context area, and suboptimal targeting has important implications. An added advantage of RMT is the graphical targeting illustration: the match between item locations and person locations (figure 1). Here, the measurement of people with mild upper limb dysfunction is limited, and the ceiling effect indicates that a notable proportion of people had no measured upper limb dysfunction post-stroke. Therefore, despite satisfying published criteria,[23 24] the ceiling effect represents a cohort of the sample for whom changes within people and differences between people will be underestimated.[37] This has implications for end point measurement and selection of outcome measures early post-stroke.

When selecting end point measures, scales should be targeted to clinical *settings* as well as the sample. Therefore, an important consideration is the nature of the items in relation to the intended context of use. For example, 16 DASH items concern activities of daily living (ADLs) which are potentially difficult for people to report meaningfully within acute settings. People may have guessed their abilities which could explain the high ability findings in a disorder where upper limb dysfunction is common. Guessing could also account for the high number of misfitting *persons* detected by RMT analyses compared to those found in MS (stroke=36%; MS=8%).[18]

It is important to consider the impact of cognitive impairment associated with stroke as this can affect peoples' responses to PROs, for example, by producing inconsistent, unlikely or random item endorsements. Such invalid responses can be reflected in the person misfit statistics. In this study, we did not assess cognitive status using formal cognitive testing to determine whether individuals might give valid responses; instead this judgement was made by a research nurse. We acknowledge this is a limitation. However, it is difficult to determine the point at which a person's responses to a PRO can be considered invalid. This area is complex and requires further investigation. Rasch measurement methods profile individual person response patterns which can allude to the presence of cognitive impairment, especially if responses are way out of keeping. However, despite notable levels of disability, person fit was better in MS (a condition in which cognition is affected) than in stroke (8% vs 36% misfit) as outlined above. This could lend support to our suggestion that misfit may have been due to people guessing their abilities within a context where they were unable to perform many of the ADLs included in the DASH.

Responsiveness analyses showed that the DASH recorded improvements in the subsample of people who had upper limb impairments (n=99; see online supplementary table S1 for sample characteristics). RMT methods are able to go beyond CTT methods of group-level responsiveness testing and examine individual person-level change (table 3). Our results are in keeping with clinical expectations of early post-stroke recovery. However, a proportion of people did get worse, some significantly. One explanation is that, during recovery, some people may have become more aware of their limitations and difficulties, and provided more accurate reporting at the 6-week follow-up. Further work is required to examine these individual's responses and help explain anomalies.

A limitation of this study is that the DASH was administered to patients admitted to the HASU regardless of whether any upper limb impairment existed or not. This is because a predetermined battery of measures was administered. However, it highlights the need for careful consideration of instrument selection and patient recruitment to ensure appropriate scales are targeted to appropriate people.

A further limitation is that we did not undertake standard validity testing. However, correlations with other measures provide circumstantial validity evidence only and would have added little to our findings.[38 39] Nevertheless, comparison of subgroups with and without upper limb impairment provides some evidence of validity: mean DASH scores between the two subsamples were significantly different implying that the DASH could discriminate between groups known to differ in level of upper limb function (see online supplementary tables S1 and S2).

One aim of our study was to compare and contrast two psychometric methods: CTT and RMT. CTT methods, based on weak measurement theory,[16 29] are limited in their ability to provide detailed diagnostic item-level and person-level

Table 4 DASH: item fit statistics ordered by location (n=271)

			Item locations		Item fit indicators	
Item	Label	Threshold ordering	Estimate	SE	Fit residual	χ^2 value
19	Recreational activities: move arm	✓	−1.61	0.10	−1.19	3.58
18	Recreational activities: force/impact	✓	−1.45	0.10	−1.42	3.03
12	Change lightbulb	✓	−0.97	0.10	**−4.40**	8.19
8	Garden	✓	−0.89	0.10	**−5.32**	13.96
11	Carry heavy object	✓	−0.81	0.10	**−4.14**	8.41
7	Do heavy chores	✓	−0.75	0.10	**−5.23**	10.63
30	Feeling less capable	✓	−0.59	0.10	**12.39**	**95.96**
10	Carry shopping bag	✓	−0.45	0.09	−2.24	8.07
1	Open new jar	✓	−0.39	0.10	**−4.65**	**18.69**
9	Make bed	✓	−0.35	0.10	**−4.86**	18.29
23	Limited in work/daily activities	✓	−0.35	0.10	−0.37	2.55
4	Prepare meal	✓	−0.34	0.10	**−4.75**	17.70
17	Recreational activities: little effort	✓	−0.27	0.10	**−4.62**	**19.57**
6	Place object on shelf	✓	−0.20	0.10	**−4.93**	15.93
22	Interference with social activities	✓	−0.16	0.09	0.97	2.90
5	Push open heavy door	✓	−0.15	0.10	**−3.57**	11.18
20	Manage transportation needs	✓	−0.13	0.09	−0.98	11.10
13	Wash hair	✓	0.01	0.10	**−3.95**	17.83
27	Weakness in arm, shoulder, hand	✓	0.06	0.10	−1.31	5.95
14	Wash back	✓	0.16	0.10	−2.33	14.75
3	Turn key	✓	0.27	0.09	2.20	11.34
16	Use knife to cut food	✓	0.30	0.10	−2.36	11.99
21	Sexual activities	✓	0.35	0.08	**6.61**	**30.81**
15	Put on sweater	✓	0.58	0.10	**−2.99**	12.41
2	Write	×	0.89	0.08	**4.55**	**153.36**
28	Stiffness in arm, shoulder, hand	✓	1.15	0.10	2.17	16.20
25	Pain performing an activity	×	1.35	0.10	0.91	**48.99**
24	Pain in arm, shoulder, hand	×	1.43	0.10	0.92	**44.23**
26	Tingling in arm, shoulder, hand	×	1.53	0.10	**2.59**	**53.42**
29	Difficulty sleeping	✓	1.78	0.11	1.74	9.80

Bold highlighted values indicate items falling outside recommended limits (fit residual −2.5 to+2.5) or value statistically significant (χ^2).
DASH, Disabilities of the Arm, Shoulder and Hand.

examinations, and there is a lack of criteria against which to make judgements on scale and item performance. The existing arbitrary criteria of traditional methods are based on assumptions that cannot be tested formally.[29 40]

The added value of RMT is highlighted by the limitations identified: disordered response categories, item misfit and suboptimal targeting. Importantly, RMT analyses provide the evidence-base for modifying and improving the DASH for future application in acute stroke research.

Future DASH research should include examinations of differential item functioning (DIF). These are sophisticated and detailed tests of the extent to which items perform differently across groups.[41 42] DIF examinations would enable, for example, comparisons of item performance across gender, age, diagnosis, treatment and disability level. However, like all RMT analyses, interpreting DIF results appropriately requires experience. Findings should be considered in light of expected clinical differences. There is a need to differentiate between real and artificial DIF, and clinically significant from clinically nonsignificant DIF. Sample size must also be considered as DIF tests are sample-size dependent. All too often investigators take a binary approach to DIF interpretation and other results generated by analysis, as they 'Rasch' their rating scale data and remove items to make the data fit the Rasch measurement model. Such post hoc data modelling approaches are not consistent with the experimental, hypothesis-testing, diagnostic RMT paradigm articulated by Rasch[17] and Andrich.[43 44]

Psychometric analysis plays a key role in scale development and testing to ensure that scales provide scientifically robust, clinically meaningful and clinically interpretable measures. The sophisticated techniques of RMT can help to ensure that PRO instruments are robust measures of the health constructs they purport to quantify.

Contributors KB, TA, AR and EDP contributed to study concept and design. KB contributed to acquisition of the data. LB and JH contributed to analysis and interpretation of the data. KB, LB, JH and EDP contributed to drafting of the manuscript. LB, JH, TA, AR and EDP contributed to critical revision of the manuscript for important intellectual content. EDP obtained funding and was involved in study supervision. JH had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors have approved the submitted version of the paper.

Funding This study was funded by the Stroke Association (grant number TSA 2007/14). EDP was supported by National Institute for Health Research University College London Hospitals Biomedical Research Centre.

Competing interests None declared.

Patient consent Obtained.

Ethics approval Ethical approval was obtained from the Joint Research Ethics Committee of the Institute of Neurology and the National Hospital for Neurology and Neurosurgery.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Nakayama H, Stig Jørgensen H, Otto Raaschou H, *et al*. Recovery of upper extremity function in stroke patients: the Copenhagen stroke study. *Arch Phys Med Rehabil* 1994;75:394–8.

2 Lawrence ES, Coshall C, Dundas R, *et al*. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke* 2001;32:1279–84.

3 Velstra IM, Ballert CS, Cieza A. A systematic literature review of outcome measures for upper extremity function using the international classification of functioning, disability, and health as reference. *PM R* 2011;3:846–60.

4 Connell LA, Tyson SF. Clinical reality of measuring upper-limb ability in neurologic conditions: a systematic review. *Arch Phys Med Rehabil* 2012;93:221–8.

5 Lang CE, Bland MD, Bailey RR, *et al*. Assessment of upper extremity impairment, function, and activity after stroke: foundations for clinical decision making. *J Hand Ther* 2013;26:104–15.

6 Mathiowetz V, Volland G, Kashman N, *et al*. Adult norms for the box and block test of manual dexterity. *Am J Occup Ther* 1985;39:386–91.

7 Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res* 1981;4:483–92.

8 Baker K, Cano SJ, Playford ED. Outcome measurement in stroke: a scale selection strategy. *Stroke* 2011;42:1787–94.

9 Department of Health. *Guidance on the routine collection of patient reported outcome measures (PROMs)*. London, UK: Department of Health, 2008.

10 Food and Drug Administration. *Patient reported outcome measures: use in medical product development to support labelling claims*. U.S. Department of Health and Human Services, 2009.

11 Duncan PW, Jorgensen HS, Wade DT. Outcome measures in acute stroke trials: a systematic review and some recommendations to improve practice. *Stroke* 2000;31:1429–38.

12 Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder, and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med* 1996;29:602–8.

13 Ring D, Kadzielski J, Fabian L, *et al*. Self-reported upper extremity health status correlates with depression. *J Bone Joint Surg Am* 2006;88:1983–8.

14 Hoang-Kim A, Pegreffi F, Moroni A, *et al*. Measuring wrist and hand function: common scales and checklists. *Injury* 2011;42:253–8.

15 Ashford S, Slade M, Malaprade F, *et al*. Evaluation of functional outcome measures for the hemiparetic upper limb: a systematic review. *J Rehabil Med* 2008;40:787–95.

16 Lord FM, Novick MR. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

17 Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Education Research, 1960.

18 Cano SJ, Barrett LE, Zajicek JP, *et al*. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. *Mult Scler* 2011;17:214–22.

19 Hobart JC, Freeman JA, Lamping DL, *et al*. The SF-36 in multiple sclerosis (MS): why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry* 2001;71:363–70.

20 Ware JE Jr, Harris WJ, Gandek B, *et al*. *MAP-R for windows: multitrait/multi-item analysis program—revised user's guide*. Boston, MA: Health Assessment Lab, 1997.

21 McHorney CA, Ware JE Jr, Lu JF, *et al*. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* 1994;32:40–66.

22 Likert RA. A technique for the measurement of attitudes. *Arch Psychol* 1932;140:5–55.

23 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995;4:293–307.

24 Holmes WC, Shea JA. Performance of a new, HIV/AIDS-targeted quality of life instrument in asymptomatic seropositive individuals. *Qual Life Res* 1997;6:561–71.

25 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.

26 Eisen M, Ware JE Jr, Donald CA, *et al*. Measuring components of children's health status. *Med Care* 1979;17:902–21.

27 Wright BD, Masters GN. *Rating scale analysis: Rasch measurement*. Chicago: MESA, 1982.

28 Andrich D. *Rasch models for measurement*. Beverly Hills, CA: Sage Publications, 1988.

29 Hobart JC, Cano SJ. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009;13.

30 Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978;43:561–73.

31 Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas* 2008;9:200–15.

32 Andrich D. An index of person separation in latent trait theory, the traditional KR20 index, and the Guttman scale response pattern. *Educ Res Perspect* 1982;9:95–104.

33 Andrich D, Sheridan B, Luo G. *RUMM2030: a windows program for the analysis of data according to Rasch Unidimensional models for measurement*. Perth, WA: RUMM Laboratory Pty Ltd, 2010.

34 Cohen J. *Statistical power analysis for the behavioural sciences*. 2nd edn. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

35 Luo G, Andrich D. Estimating parameters in the Rasch model in the presence of null categories. *J Appl Meas* 2005;6:128–46.

36 Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* 2007;46:1–18.

37 Barrett LE, Cano SJ, Zajicek JP, *et al*. Can the ABILHAND handle manual ability in MS? *Mult Scler* 2013;19:806–15.

38 Hobart J, Cano S, Baron R, *et al*. Achieving valid patient-reported outcomes measurement: a lesson from fatigue in multiple sclerosis. *Mult Scler* 2013;19:1773–83.

39 Stenner AJ, Smith M, Burdick D. Towards a theory of construct definition. *J Educ Meas* 1983;20:305–16.

40 Novick MR. The axioms and principal results of classical test theory. *J Math Psychol* 1966;3:1–18.

41 Hagquist C, Andrich D. Measuring subjective health among adolescents in Sweden. *Soc Indic Res* 2004;68:201–20.

42 Andrich D, Hagquist C. Real and artificial differential item functioning. *J Educ Behav Stat* 2012;37:387–416.

43 Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res* 2011;11:571–85.

44 Andrich D. The legacies of R. A. Fisher and K. Pearson in the application of the polytomous Rasch model for assessing the empirical ordering of categories. *Educ Psychol Meas* 2013;73:553–80.