Measurement properties of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ): a study based on Classical Test Theory

Lisette M. van Leeuwen, Lidwine B. Mokkink, Christian P. Kamm, Vincent de Groot, Pauline van den Berg, Raymond W. J. G. Ostelo & Bernard M. J. Uitdehaag

Taylor & Francis
Taylor & Francis Group

ASSESSMENT PROCEDURES

 OPEN ACCESS

Measurement properties of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ): a study based on Classical Test Theory

Lisette M. van Leeuwen[a], Lidwine B. Mokkink[b], Christian P. Kamm[c,d], Vincent de Groot[e], Pauline van den Berg[f], Raymond W. J. G. Ostelo[b,g] and Bernard M. J. Uitdehaag[d]

[a]Department of Otolaryngology – Head and Neck Surgery, Section Ear & Hearing, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands; [b]Department of Epidemiology and Biostatistics, EMGO institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands; [c]Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland; [d]Department of Neurology, VU University Medical Center, Amsterdam, The Netherlands; [e]Department of Rehabilitation Medicine, VU University Medical Center, Amsterdam, The Netherlands; [f]Nieuw Unicum Zandvoort, Zandvoort, The Netherlands; [g]Department of Health Sciences, Faculty of Earth and Life Science, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands

ABSTRACT

Purpose: The construct validity, test–retest reliability, and measurement error of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ) were examined. Additionally, the influence of administration-method on reliability and measurement error was investigated.

Method: 112 Dutch adult MS-patients from an academic- and a residential care-facility participated. Questionnaires were administered on paper, online or as interview, and patients performed several performance tests. Construct validity was assessed by testing pre-defined hypotheses. Reliability was assessed using Intraclass Correlation Coefficients (ICCs), Standard Error of Measurements (SEMs) and Smallest Detectable Changes (SDCs).

Results: For construct validity ($N = 105$) 9 of 13 hypotheses were confirmed (69%). As expected, the AMSQ showed moderate to strong relationships with the instruments measuring similar constructs. The test–retest reliability coefficient was 0.96 (95% Confidence Interval 0.94–0.97); SEM was 6.3 (6.3% of scale range); SDC was 17.5 (on a sale from 0 to 100). Different administration-methods showed good reliability (ICC 0.88–0.94) and small standard errors (SEM 5.6–7.2).

Conclusion: The AMSQ shows satisfying results for validity and excellent reliability; allowing for proper use in research. Due to a large SDC value, caution is needed when using the AMSQ in individual patient care. Further research should determine whether the SDC is smaller than the minimal important change.

➤ IMPLICATIONS FOR REHABILITATION

- The Arm Function in Multiple Sclerosis Questionnaire (AMSQ) measures activity limitations due to hand and arm functioning in patients with Multiple Sclerosis (MS).
- Results of this study confirm adequate validity and reliability of the AMSQ in patient with MS.
- The equivalence of scores from online, paper or interview administration is supported.
- A change score of ≥18 points on the scale of the AMSQ (on a scale 0–100) needs to occur to be certain a change beyond measurement error has occurred in an individual patient.

Introduction

Multiple Sclerosis (MS) is a progressive disease and is highly associated with increased physical disability. Limitations in hand and arm functioning are present in up to 76% of patients [1–4] including patients with low disease severity.[3] Impairments may include tremor, coordination deficit and muscle weakness.[5] Limitations in hand and arm function have significant negative implications on activities of daily living (ADL) (e.g., eating, dressing, grooming),[6] living independently, and quality of life (QoL),[7–9] and are associated with high societal costs due to loss of work and high direct costs due to utilization of care.[10] Given their impact, valid assessment of hand and arm functioning is important in both clinical practice and clinical trials, i.e., for the comprehensive

evaluation of therapeutic effectiveness and in the development of treatment strategies.[11,12]

Patient Reported Outcome Measures (PROMs) capture the patient's perspective of a person's health condition and are advised when measuring functioning.[13,14] Moreover, along with the recognition of the importance of patient-centered care, PROMs have gained increased importance in the assessment of MS. There are several PROMs for measuring upper extremity function.[15–19] However, no unidimensional disease specific PROM is available for measuring arm and hand functioning in persons with MS. Therefore, the Arm Function in Multiple Sclerosis Questionnaire (AMSQ) was recently developed [20] to measure activity limitations due to hand and arm functioning in patients with MS. The term "activity limitations" was defined according to

---

the International Classification of Functioning, Disability and Health (ICF) as any difficulties an individual may have in executing activities.[21] The AMSQ was based on a literature review and with involvement of experts and patients, and using item response theory (IRT) methods.[20] More details are published elsewhere.[20] In total, 31 items were included in the questionnaire constituting a unidimensional scale. All items are formulated as "during the past two weeks, to what extent has MS limited your ability to …". Response categories are *not at all*, *a little*, *moderately*, *quite a lot*, *extremely*, and *no longer able to*. The aim is to develop a computer adaptive test in the future, requiring an item bank containing a sufficient number of items that cover the range of activity limitations due to hand and arm functioning for patients with MS. In the present study, we further investigated the quality of the AMSQ using classical test theory (CTT) methods. The aim of this study was to investigate construct validity, test–retest reliability, and measurement error. Often in research, different modes of administration are used, such as self-report, or interview. Moreover, a self-report version can either be administered at the clinic or at home, by paper and pencil or online. In general, different ways of administration can influence the scores. Therefore, we subsequently investigated whether the mode of administration influenced the reliability and measurement error of the AMSQ.

Method

Study design and patients

In this test–retest design, a prospective cohort of patients with MS was recruited at the VU University Medical Center (VUmc) in Amsterdam and the residential and facility center for physically handicapped, Nieuw Unicum (NU) in Zandvoort, both in the Netherlands. A sample of patients was enrolled at the VUmc at regular patient visits at the outpatient clinic or in the context of clinical ongoing research projects. In addition, patients were recruited from advertisements on Dutch websites, i.e., www.msweb.nl and www.msvamsterdam.nl. Patients at NU were recruited by the staff of the center. Inclusion criteria were: self-indicated diagnosis of any type of MS, age above 18 years, and adequate understanding of the Dutch language. Patients with clinically observable severe cognitive impairments were excluded. The study was approved by the Medical Ethical Committee of the VU University Medical Center, Amsterdam; the Netherlands (reference number 2012/296). Data collection was carried out between November 2012 and June 2013.

Procedures

Patients were examined at VUmc or at NU. All patients were asked to sign an informed consent form, and to complete a questionnaire, containing demographic variables and disease specific questions (i.e., age, gender, disease duration, and MS type), a self-administered version of the Expanded Disability Status Scale (EDSS),[22] and several other PROMS including the AMSQ (see below) provided online or on paper. Furthermore, all patients were interviewed to assess level of disability due to MS (i.e., The Guy's Neurological Disability Scale, see below) and were asked to perform several performance tests (see below). At NU, the questionnaires were administered as interview, as most patients were unable to fill out the questionnaires themselves due to hand and arm impairments. At VUmc, patients self-administered the questionnaires.

After two to four weeks we asked patients again to complete the AMSQ, assuming that this time interval was sufficient to minimize recall bias, yet short enough for their hand and arm function to remain unchanged. To check whether patients were stable in the meantime, a Global Perceived Effect (GPE) scale about perceived changes in hand and arm functioning was administered (see below).

Measurements

Patient reported outcome measures

- *The AMSQ* measures activity limitations due to hand and arm function in patients with MS. All items fitted into the graded response model, which is an IRT model, and no differential item functioning (DIF) was found for the variables type of MS, gender, administration version, and test length. IRT based reliability was 0.95.[20] The items of the AMSQ are provided in Appendix 1. As this study is based on CTT, we calculated sum scores (range 0–100) instead of IRT-based trait level (θ) scores. Higher scores indicate more limitations in hand and arm function. No scores were imputed, only complete cases were used.

- *The Guy's Neurological Disability Scale (GNDS)* is an structured interview assessing level of disability (i.e., activity limitations).[23,24] The GNDS contains 12 functional domains, but we used only the total GNDS score (range 0–60) and the upper limb disability sub score (range 0–5). Each domain contains four to ten dichotomous items. Based on the given answers, domain scores are ascribed on a six-point severity scale, e.g., the upper limb disability scoring ranges from *no upper limb problem* (score = 0) to *unable to use either arm for any purposeful movements* (score = 5). No missing variables were expected, as the scale was administered as interview.

- *The RAND-36* is a license free version of the Short-Form Health Survey-36 (SF-36),[25,26] and comprises 36 items assigned to eight scales: physical functioning (10 items), role-physical (4 items), bodily pain (2 items) and general health (5 items), vitality (4 items), social functioning (2 items), role-emotional (3 items) and mental health (5 items). The scores of each domain were transformed so that a higher score indicates better health status (range 0–100).

- *The Multiple Sclerosis Impact Scale-29 (MSIS-29)* measures the impact of MS on daily living, comprising a physical impact scale (MSIS-29 physical) and a psychological impact scale (MSIS-29 psychological).[27,28] All items have a five-point Likert scale ranging from *not at all* to *extremely*, with higher scores indicating higher impact (range physical subscale 20–100; range psychological subscale 9–45).

- *The Multiple Sclerosis Impact Profile (MSIP)* measures disability in patients with MS. Only the subscale measuring activities of daily living was utilized.[29] Answering is on a four-point scale, ascending in degree of need for supporting tools and/or help from others, i.e., higher scores indicated more dependence (range 7–28).

- GPE scale.[30] At follow-up patients were asked *How would you rate your hand/arm functioning, compared to two weeks ago?*. Response options were: (1) *much better than two weeks ago*, (2) *somewhat better than two weeks ago*, (3) *about the same as two weeks ago*, (4) *somewhat worse than two weeks ago*, and (5) *much worse than two weeks ago*. Patients who reported *much better* or *much worse* on the GPE scale were regarded as unstable patients and were excluded from reliability analyses.

Missing items on the RAND-36, MSIS-29 and MSIP were imputed as recommended with patient-specific mean values of completed items.

Performance tests

All administered performance-based tests are designed to measure each hand or arm independently. Scores were obtained for both hands/arms. Because the AMSQ was designed regardless of hand-dominance, we averaged the scores of the dominant and non-dominant hand for each performance based test.

- *The Action Research Arm test (ARAT)* was used to measure fine and gross motor dexterity.[31] The test consists of five subtests: grasp (6 items), grip (4 items), pinch gross (4 items), pinch fine (4 items), and gross movement (3 items), comprising a total of 19 movements to be performed by the patient. Each movement is scored on a scale from *no movement possible* (score = 0) to *normal movement* (score = 3) (range 0–57).
- *The Nine Hole Peg Test (9-HPT)* was used to measure upper extremity function,[32] and involves placing and removing nine pegs in a pegboard. The time to perform the test was measured.
- *The Coin Rotation Task (CRT)* was used to measure fine motor dexterity of the hands,[33] and involves rotating a US five-cent coin as fast as possible using the thumb, index and middle fingers. The time to perform 20 half-turns was measured.
- *The hand held JAMAR dynamometer* was used to measure isometric grip strength of the hand.[34] The test was performed following standardized instructions and positioning recommended by the American Society of Hand Therapists.[35]
- *The Modified Ashworth Scale (MAS)* was used for measuring muscle spasticity,[36] by measuring resistance to passive movement about the elbow joint on a 6-point scale from *no increase in tone* to *limb rigid in flexion or extension* (range 0–5). MAS scores were dichotomized into no spasticity (i.e., MAS = 0) and spasticity (i.e., MAS $\geq$ 1).

Except for the ARA test and MAS, all performance tests were administered twice on both hands and the best value was taken as score for each hand. If a patient could not perform a timed-test due to hand and arm impairments, a maximum value of 300 seconds was used (according to the manual of the NHPT and used for other tests [37]).

Statistical analyses

We produced descriptive statistics (means, medians, and SDs) for the scores of the measurements, and investigated the frequencies of missing data.

Construct validity

Construct validity was assessed by the degree to which the sum scores of the AMSQ were consistent with predefined hypotheses regarding relationships between the AMSQ and the other measures. We formulated 13 hypotheses presented in Table 1. Moderate to high correlations are expected between the AMSQ and other PROMs measuring physical functioning (hypotheses 1–6). Low correlations are expected between the AMSQ and other PROMs measuring different constructs (hypotheses 7, 8). Moderate to high correlations are expected between the AMSQ and all performance tests, as they all assess aspects of upper limb functioning. Though, a hierarchy in strength of the linear relationship between the AMSQ and the different performance measures was expected (hypotheses 9–12), i.e., the ARAT reflects the same construct as the AMSQ, i.e., "hand/arm functioning", and therefore the strongest correlations coefficient was expected. The 9-HPT, CRT and JAMAR hand strength dynamometer measure narrower constructs compared to the construct measured by the AMSQ, and therefore, lower correlation coefficients were expected. In addition, one hypothesis regarding expected differences in AMSQ mean sum scores in patients with spasm and patients without spasm was defined (hypothesis 13).

Spearman's rho correlations were used for assessing all hypothesized relations between the AMSQ and PROMs and performance-based measures because scores were non-normally distributed. Correlation was considered as low <0.30; moderate 0.30–0.59; and high $\geq$0.60.[38] Group comparison (patients with spasm versus patients without spasm) was made by a Mann–Whitney U Test with a *p* cutoff value of 0.05.

Reliability

To investigate test–retest reliability of the AMSQ, we calculated one-way ICC for the whole sample due to an incomplete design [39] (Reliability question 1). The questionnaires were administered online, on paper, or as an interview. Mode of self-administration could vary between baseline and retest measurement (i.e., a patient filled out the baseline questionnaire on paper at the VUmc, and completed the retest online at home). For patients at NU baseline AMSQ was administered by two researchers (L.v.L. or L.M.), and follow-up administration was performed by one of the eight physiotherapists following a time interval of two to four weeks after initial assessment. In addition, we investigated whether there was a systematic difference between two measurements due to differences in mode of administration (Reliability question 2), and whether there was a systematic difference between two measurements due to different observers (Reliability question 3). These two questions were investigated by calculating ICC two-way ANOVA random effect models for agreement for patients who completed baseline and retest questionnaires in different ways of administration (i.e., paper at baseline and online at follow-up), and for patients who were interviewed. An ICC value of 0.70, in a sample of 50 patients was recommended as a minimum standard for reliability.[40]

Measurement error

The measurement error was determined by calculating the standard error of measurement (SEM),[41] i.e., the square root of the error variance from the ICC formula. In addition, measurement error was expressed as smallest detectable change (SDC). The SDC represents the minimal change that a patients must show on the scale to ensure (with 95% confidence) that the observed change is real and not just measurement error. The SDC was calculated at a 95% confidence interval by multiplying the SEM by 1.96 and by the square root of 2.[30] All statistical analyses were performed using the Statistical Package for Social Science (SPSS) version 20.0 (Chicago, IL).

Results

Patient characteristics and response rate

Patient characteristics are presented in Table 2. In total 112 patients with MS participated, of which 77 patients were

Table 1. Specific hypotheses and correlation coefficients of the AMSQ with other measurement instruments ($N = 105$).

	Hypothesis	Confirmed yes/no	Result (r)
PROMs			
1	A high correlation was expected between the AMSQ and GNDS upper limb disability domain, because they measure the same construct based on the framework of the ICF.	Yes	0.84
2	A lower correlation was expected between the AMSQ and the total sum score of the GNDS, as compared with the correlation between the AMSQ and the GNDS upper limb disability domain.	Yes	0.69 vs 0.84
3	A moderate correlation between the AMSQ and MSIS-29 physical subscale was expected, because the MSIS-29 physical subscale measures physical impact of MS in daily life, and is not specifically focusing on hand and arm functioning.	No	0.77[a]
4	A higher correlation between the AMSQ and MSIS-29 physical subscale was expected as compared with the correlation between the AMSQ and RAND-36 physical functioning subscale, because the MSIS-29 is a MS specific instrument and the RAND-36 is not.	No	0.77 vs −0.81[b]
5	A higher correlation between the AMSQ and the physical functioning subscale of the RAND-36 was expected as compared with RAND-36 physical component summary score.	Yes	−0.81 vs −0.61[b]
6	A higher correlation between the AMSQ and the MSIP-ADL subscale was expected as compared with the MSIS-29 physical subscale, because limitations in upper extremity functioning strongly influence ADL and the AMSQ focuses on functioning of ADL.	Yes	0.80[c] vs 0.77
7	Low correlations were expected between the AMSQ and the MSIS-29 psychological subscale, because they measure non-similar constructs.	Yes	0.17[a]
8	Low correlations were expected between the AMSQ and RAND-36 mental component summary score (MCS), because they measure non-similar constructs.	Yes	0.02[b]
Performance tests[d]			
9	High correlations were expected between the AMSQ and ARA test, because the ARAT is developed for upper extremity motor function including both grip and pinch, and fine and gross arm movements.	Yes	−0.65[e]
10	Somewhat lower correlations were expected between the AMSQ and 9-HPT, because in contrast to the ARAT, the 9HPT focuses more specifically on hand function and finger dexterity while the AMSQ assesses the whole upper extremity.	No	0.77[f] vs −0.65
11	It was expected that the AMSQ will show somewhat lower correlations with the CRT as compared with the correlation between the AMSQ and 9-HPT, because the CRT focuses even more specifically on the hand and finger dexterity as compared with the 9-HPT.	No	0.77[g] vs 0.77
12	It was expected that the AMSQ will show lower correlations with the JAMAR hand strength dynamometer, as compared with the correlations between the AMSQ and the CRT, because the AMSQ does not specifically measures hand strength.	Yes	−0.67 vs 0.77
13	It was expected that patients with spasticity (MAS ≥1) score statistically significantly higher on the AMSQ as compared with patients without spasticity (MAS =0), because spasticity is a causative for activity limitations due to impairments of the arm and hand.	Yes	$p < 0.05$

Low = correlation <0.30; Moderate = correlation 0.30–0.59; High = correlation ≥0.60.
r: Spearman's rho correlation; PROMs: Patient Reported Outcome Measures; AMSQ: Arm function in Multiple Sclerosis Questionnaire; GNDS: Guy's Neurological Disability Scale; Rand-36: The 36 item short form health survey; PF: physical functioning; MSIS-29: Multiple Sclerosis Impact Scale; MSIP-ADL: The Multiple Sclerosis Impact Profile-Activities of Daily Living; 9HPT: Nine Hole Peg Test; CRT: Coin Rotation Task; ARAT: Action Research Arm Test; JAMAR: dynamometer for measuring hand strength.
[a]Six items were imputed.
[b]$N = 95$; 10 patients did not fill in the RAND-36. Nine items were imputed.
[c]One item was imputed.
[d]Averaged performance test scores for $N = 105$, dominant/non-dominant hand scores for $N = 104$ due to 1 missing item on hand dominance.
[e]$N = 91$; 14 patients had no scores, because they could not perform the test due to practical problems.
[f]13 and 16 patients had a value of 300 seconds on dominant and non-dominant hand, respectively, because they could not perform the test due to hand and arm impairments.
[g]26 and 28 patients had a value of 300 seconds on dominant and non-dominant hand, respectively, because they could not perform the test due to hand and arm impairments.

Table 2. Patient characteristics ($N = 112$).

Age (years)	Mean (SD)	47.6 (11.6)
	min-max	20–72
Gender	female	55.4%
Recruitment method	Active application at VUmc and from websites	11.6%
	Regular patient visits at VUmc	48.2%
	Ongoing research projects at VUmc	8.9%
	Inpatient recruitment at NU	31.3%
Disease duration (years)	median	13
	min-max	1–49
Self. Adm. EDSS	mean (SD)	5.4 (2.7)
	min-max	0–9
MS type	Relapsing Remitting (RR)	50.5%
	Secondary Progressive (SP)	21.1%
	Primary Progressive (PP)	11.9%
	CIS and unknown	16.5%

SD: standard deviation, min: minimum; max: maximum; Self. Adm.: self-administered; EDSS: Expanded Disability Status Scale; CIS: Clinically Isolated Syndrome. Missing values were found in the variables age ($N = 1$), disease duration ($N = 8$) and MS type ($N = 3$).

recruited at the VUmc and 35 patients at NU. All subjects from NU were residential. Sum scores on the AMSQ were only calculated for patients who completed all 31 items (94%). Five patients had one missing item and two patients had two or three missing items. The total of missing items of the AMSQ was less than 1%. Mean sum score of the AMSQ was 27.8 (SD = 31.8) and the median was 12.9 (range 0–100). For the validity analyses we used the scores of the 105 patients with complete cases on the AMSQ. With regard to the follow-up measurement, 14 of the 77 patients that used self-administration did not complete the follow-up questionnaire (response rate of 82%), and four patients did not fully complete the AMSQ. All 35 patients who were interviewed completed the follow-up measurement, but three patients had missing items on the AMSQ. Taken together, 91 patients were eligible for reliability analyses.

Table 3. Reliability question numbers and reliability of the AMSQ.

nr	Method of administration Baseline	Re-test	N	ICC method	ICC (95%CI)	σ_p^2	σ_o^2	$\sigma_{residual}^2$	SEM (% of scale range)	SDC	d
1	Paper Interview Online	Online Interview Online	86	One-way ANOVA	0.96 (0.94–0.97)	955.23	NA	38.97	6.3 (6.3%)	17.5	−0.20
2	Paper	Online	43	Two-way ANOVA	0.88 (0.79–0.93)	242.36	0.97	31.34	5.6 (5.6%)	15.6	−1.85
3	Interview	Interview	31	Two-way ANOVA	0.94 (0.88–0.97)	796.98	0.00	51.62	7.2 (7.2%)	20.0	1.02

ICC: intraclass correlation coefficient; CI: confidence interval; SEM: standard error of measurement; SDC: smallest detectable change; NA: not applicable; σ_p^2 the variance of the patients (i.e., the systematic differences between the "true" scores of the patients); σ_o^2 the variance due to systematic differences between the measurements/observers; $\sigma_{residual}^2$ the random error variance; d: the systematic difference.

Construct validity

The correlation coefficients between the AMSQ mean sum score and the mean values of other measures are presented in Table 1. In summary, 9 out of the 13 predicted hypotheses were confirmed (69.2%). As expected, moderate to high correlation coefficients were found between the AMSQ and (sub) scales measuring physical functioning (hypotheses 1–6) and low correlation coefficients were found between the AMSQ and (sub) scales measuring non-similar constructs (hypotheses 7 and 8). The correlations between the AMSQ and all performance-based hand and arm function tests (hypotheses 9–12) were moderate to high, as expected. Although the hierarchical order of the three comparator tests was not as expected, i.e., the ARAT showed a lower correlation coefficient with the AMSQ when compared with the observed coefficients of the other performance tests. In line with expectations, we found a significant difference between patients with spasticity ($N = 30$) versus no spasticity ($N = 71$) on the AMSQ sum score (U $= 557.5$; $p < 0.05$).

Reliability

Five patients reported "much better" or "much worse" on the GPE and were excluded for analyses. Of the remaining sample ($N = 86$), 55 patients self-administered the questionnaires and 31 patients were interviewed at baseline and follow up. Of the patients who self-administered the questionnaires, 43 patients used different modes of administration for baseline and follow-up measurement (i.e., paper at baseline and online at follow-up), and 12 patients used the same method (i.e., both administrations were online). None of the patients self-administered the questionnaire twice on paper. The results addressing the three research questions concerning reliability are presented in Table 3.

Discussion

The AMSQ is a newly developed tool to measure activity limitations due to hand and arm functioning in patients with MS. The first quality assessment by IRT methods showed good results, and in this second study we evaluated the psychometric properties of the AMSQ using traditional CTT methods. We assessed construct validity, reliability and measurement error.

The validity analysis showed satisfying results for construct validity of the AMSQ (confirmed hypotheses 69.2%). As expected, the AMSQ showed moderate to strong linear relationships with PROMs measuring similar constructs and the performance-based tests. Hypotheses 3, 4, 10 and 11 were not confirmed. Surprisingly, a stronger relationship was found between the AMSQ and RAND-36 physical subscale (which is developed for the general population) when compared with the relationship between the AMSQ and MSIS-29 physical subscale, which similar to the

AMSQ, was specifically developed for patients with MS (i.e., hypotheses 3 and 4). The items of the AMSQ and RAND-36 all ask only about limitations in performing activities, while 7 out of 20 items of the MSIS-29 physical subscale ask about limitations due to specific causes, such as balance, clumsy, stiffness, tremor, and spasms. It is likely that therefore the construct measured by the MSIS-29 is different from the constructs as measured by the AMSQ and the RAND-36. Furthermore, the correlations between the AMSQ and the ARAT in our study were not as expected (i.e., hypotheses 10 and 11). Originally, the ARAT was developed to assess upper extremity function following cortical injury. Although the ARAT was shown valid for measuring upper extremity functioning patients with MS,[42] the test seemed not suitable for measuring hand and arm functioning in the present study sample. One explanation might be that patients scored mostly the best possible score (i.e., 66% had an average score >55) despite claiming dexterous difficulties, as well as that in more severely disabled patients (EDSS ≥ 8) the test could not be administered due to practical problems (22%), e.g., wheelchair dependent patients could not reach all parts of the ARAT box. Similar results and remarks regarding validity and ceiling effects on the ARAT in MS patients have been reported in other studies.

The high-reliability coefficients and low measurement error support the value of the AMSQ in clinical trials with relatively small sample sizes, regardless of different modes of administration both between and within patients. The reliability coefficient determined on different modes of administration provided good support for the equivalence of scores from online, paper or interview administration. This corresponds to findings of our previous study,[20] which showed no DIF on any of the 31 items of the AMSQ for mode of administration. In clinical practice a SEM of (approximately) 6 points means that when a patient gets a score of for example 28 points, in reality the score will lie somewhere between 22 points and 34 points. The SDC was 17.5 points (on a scale of 0 to 100). This means that when an individual patient is measured over time, a change score of at least 18 points needs to occur in order to conclude that in reality (with a 95% certainty) a change beyond measurement error has occurred in an individual patient. The SDC for self-report was smaller than the SDC for interview (15.6 vs. 20.0). This could be taken into account when using the instrument, for example in a randomized controlled trial. It could also be possible that the measurement error for patients who have lower functioning level is larger, which were the patients that were interviewed. For the interpretation of change scores of PROMs, results on both the SDC and the minimal important change are needed (MIC; i.e., the smallest change in score which patients perceive as important). An instrument is useful in clinical practice if the SDC is smaller than the minimal important change. A next step therefore is to determine whether the smallest detectable change is sufficiently small, i.e., smaller than the minimal important change. Note that when using the

instrument in a study to measure change in a group of patients, the measurement error of the mean change score is much lower ($SDC/\sqrt{n}$).[41]

Since the AMSQ is developed using IRT methods, trait level (θ) scores can be obtained. The considerable advantage of theta scores is their ability to handle missing data. The analyses were repeated using trait level (θ) scores. The correlation between the sum scores and theta scores was 0.92. Similar results were obtained for construct validity as well as for reliability analyses (data not shown).

Study strengths and limitations

According to international guidelines, a minimum of 50 patients is considered adequate for assessing measurement properties.[43] We included 102 patients for validity and 86 patients for reliability analyses and thus largely met this criterion. Furthermore, we had little missing data on the AMSQ and all other questionnaires. Therefore, we do not expect that the missing data has led to bias. Some limitations need to be addressed regarding the present study. Patients self-indicated their diagnosis of MS. Because of ethical and regulatory restrictions, the diagnosis of MS was not confirmed by accessing medical records. This limitation might limit the generalizability of the results. However, given the study sample largely consisted of patients that either visited the academic hospital (VUmc) for treatment or were admitted to a residential care facility that is specialized in care for MS patients (NU), the vast majority of patients were known to us with the right diagnosis (MS). We are therefore confident the self-indicated diagnosis was valid and that this has not led to bias. Regarding the performance-based tests, missing data were mostly obtained due to practical problems using the ARAT, which could have led to bias. Furthermore, we averaged the scores of the dominant and non-dominant hands on the performance-based tests. This could have introduced bias on the representativeness of the found correlations. However, the outcome did not change using only scores for the dominant and non-dominant hand, respectively (data not shown). Another limitation was that we had loss to follow-up for the second questionnaire (13%). However, the patient characteristics of non-responders were not different as compared with responders, dismissing that this may have caused different results. Furthermore, we used a GPE to define stable patients for reliability assessment. Such a measure has several limitations,[44–46] including questionable validity, recall bias and influence of current status. Although we cannot rule out recall bias, we believe clinically important changes were unlikely to occur in two to four weeks. Moreover, the negligible systematic differences that were found might be an indication that biological variance or recall bias had no influence. Unfortunately one item of the MSIP was inadvertently not included in this study. This might have led to bias regarding the correlation coefficient for the relation between the scores on the AMSQ and MSIP.

Conclusion and practical implication

The results of this study show satisfying results for validity and excellent results for reliability in a sample of Dutch patients with MS. This second evaluation of measurement properties support that the AMSQ is an adequate scale for measuring arm and hand functioning in patients with MS in clinical research. Further research will determine whether the same results apply to translations into other languages. An English version is currently under investigation in an Irish population,[47] and another study is ongoing for translating and validating the German version of the AMSQ.[48]

References

[1] Goodkin DE, Hertsgaard D, Seminary J. Upper extremity function in multiple sclerosis: improving assessment sensitivity with box-and-block and nine-hole peg tests. Arch Phys Med Rehabil. 1988;69:850–854.

[2] Einarsson U, Gottberg K, Von Koch L, et al. Cognitive and motor function in people with multiple sclerosis in Stockholm County. Mult Scler. 2006;12:340–353.

[3] Johansson S, Ytterberg C, Claesson IM, et al. High concurrent presence of disability in multiple sclerosis. Associations with perceived health. J Neurol. 2007;254:767–773.

[4] Kamm CP, Heldner MR, Vanbellingen T, et al. Limb apraxia in multiple sclerosis: prevalence and impact on manual dexterity and activities of daily living. Arch Phys Med Rehabil. 2012;93:1081–1085.

[5] McDonald I, Compston A. The symptoms and signs of multiple sclerosis. In: Compston A, Ebers G, Lassmann H, editors. McAlpine's multiple sclerosis. 4th ed. London: Curchill Livingstone; 2006. pp. 287–346.

[6] Yozbatiran N, Baskurt F, Baskurt Z, et al. Motor assessment of upper extremity function and its relation wiht fatigue, cognitive function and quality of life in multiple sclerosis patients. J Neurol Sci. 2006;246:117–122.

[7] Månnsson E, Lexell J. Performance of activities of daily living in multiple sclerosis. Disabil Rehabil. 2004;26:576–585.

[8] Krishnan V, Jaric S. Hand function in multiple sclerosis: force coordination in manipulation tasks. Clin Neurophysiol. 2008;119:2274–2281.

[9] Sehanovic A, Dostovic Z, Smajlovic D, et al. Quality of life in patients suffering from Parkinson's disease and multiple sclerosis. Med Arh. 2011;65:291–294.

[10] Abbas D, Gehanno JF, Caillard JF, et al. Characteristics of patients suffering from multiple sclerosis according to professional situation. Ann Réadapt Méd Phys. 2008;51:386–393.

[11] Riazi A. Patient-reported outcome measures in multiple sclerosis. Int MS J. 2006;13:92–99.

[12] Cella D, Nowinski C, Peterman A, et al. The neurology quality-of-life measurement initiative. Arch Phys Med Rehabil. 2011;92:S28–S36.

[13] Department of Health and Human Services, Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), and Center for Devices and Radiological Health (CDRH). 2007. Guidance for Industry patient-reported outcome measures: use in medical product development to support labeling claims; [cited 2016 June 01] Available from: http://www.fda.gov/cder/guidance/5460dft.pdf

[14] European Medicines Agency. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medical products; 2009; [cited 2016 June 01]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf

[15] Bellamy N, Campbell J, Haraoui B, et al. Clinimetric properties of the AUSCAN osteoarthritis hand index: an evaluation of reliability, validity and responsiveness. Osteoarthr Cartil. 2002;10:863–869.

[16] Bellamy N, Campbell J, Haraoui B, et al. Dimensionality and clinical importance of pain and disability in hand osteoarthritis: development of the Australian/Canadian (AUSCAN) osteoarthritis hand index. Osteoarthritis Cartilage. 2002;10: 855–862.

[17] Penta M, Thonnard JL, Tesio L. ABILHAND: a Rasch-built measure of manual ability. Arch Phys Med Rehabil. 1998;79:1038–1042.

[18] Levine DW, Simmons BP, Koris MJ, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. J Bone Joint Surg Am. 1993;75:1585–1592.

[19] Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand. Am J Ind Med. 1996; 29:602–608.

[20] Mokkink LB, Knol DL, Van der Linden FAH, et al. The arm function in multiple sclerosis questionnaire (AMSQ): development and validation of a new tool using IRT methods. Disabil Rehabil. 2015;37:2445–2451.

[21] WHO. International classification of functioning, disability and health. Geneva: World Health Organization; 2001.

[22] Kurtzke JF. Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS). Neurology. 1983;33:1444.

[23] Sharrack B, Hughes RA. The Guy's neurological disability scale (GNDS): a new disability measure for multiple sclerosis. Mult Scler. 1999;5:223–233.

[24] Mokkink LB, Knol DL, Uitdehaag BM. Factor structure of Guy's neurological disability scale in a sample of Dutch patients with multiple sclerosis. Mult Scler. 2011;17: 1498–1503.

[25] Ware JE, Gandek B. Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project. J Clin Epidemiol. 1998;51:903–912.

[26] Aaronson NK, Muller M, Cohen PD, et al. Translation, validation, and norming of the Dutch language version of the SF-36 health survey in community and chronic disease populations. J Clin Epidemiol. 1998;51:1055–1068.

[27] Riazi A, Hobart JC, Lamping DL, et al. Multiple sclerosis impact scale (MSIS-29): reliability and validity in hospital based samples. J Neurol Neurosurg Psychiatry. 2002;73: 701–704.

[28] Hoogervorst EK, Zwemmer JN, Jelles B, et al. Multiple sclerosis impact scale (MSIS-29): relation to established measures of impairment and disability. Mult Scler. 2004; 10:569–574.

[29] Wynia K, Middel B, Van Dijk JP, et al. The multiple sclerosis impact profile (MSIP). Development and testing psychometric properties of an ICF-based health measure. Disabil Rehabil. 2008;30:261–274.

[30] de Vet HC, Terwee CB, Mokkink LB, et al. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.

[31] Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res. 1981;4:483–492.

[32] Kellor M, Frost J, Silberberg N, et al. Hand strength and dexterity. Am J Occup Ther. 1971;25:77–83.

[33] Mendoza JE, Apostolos GT, Humphreys JD, et al. Coin rotation task (CRT): a new test of motor dexterity. Arch Clin Neuropsychol. 2009;24:287–292.

[34] Paltamaa J, West H, Sarasoja T, et al. Reliability of physical functioning measures in ambulatory subjects with MS. Physiother Res Int. 2005;10:93–109.

[35] Fess EE. In: Casanova JS, editor. Clinical assessment recommendations. 2nd ed. Chicago: American Society of Hand Therapists; 1992. pp. 41–45. Grip strength.

[36] Asworth B. Preliminary trail of carisoprodol in multiples sclerosis. Practitioner. 1964;192:540–542.

[37] Fisher JS, Rudick RA, Cutter GR, et al. The multiple sclerosis functional composite measure (MSFC): an integrated approach to MS clinical outcome assessment. Mult Scler. 1999; 5:244–250.

[38] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: L. Erblaum Associates; 1988. pp. 567.

[39] Euser AM, le Cessie S, Dutch POPS-19 Collaborative Study Group, et al. Reliability studies can be designed more efficiently by using variance components estimates from different sources. J Clin Epidemiol. 2007;60:1010–1014.

[40] Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34–42.

[41] Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010; 63:737–745.

[42] Platz T, Pinkowski C, van Wijck F, et al. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer test, action research arm test and box and block test: a multicentre study. Clin Rehabil. 2005;19:404–411.

[43] Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21:651–657.

[44] Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50: 869–879.

[45] Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. J Clin Epidemiol. 2002;55:900–908.

[46] Kamper SJ, Ostelo RW, Knol DL, et al. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol. 2010;63:760–766.

[47] Chen CC, Kasven N, Karpatkin HI, et al. Hand strength and perceived manual ability among patients with multiple sclerosis. Arch Phys Med Rehabil. 2007;88:794–797.

[48] Lamers I, Kerkhofs L, Raats J, et al. Perceived and actual arm performance in multiple sclerosis: relationship with clinical tests according to hand dominance. Mult Scler J. 2013;19:1341–1348.

Appendix 1

Arm Function in Multiple Sclerosis Questionnaire (AMSQ)

Copyright of the AMSQ is held by the MS Center Amsterdam of VUmc. For translations in other languages please contact the corresponding author.

Please note that the AMSQ was developed in Dutch, and afterwards translated into English. All analyses described in this article are based on the Dutch 31-item version of the AMSQ.

Please read the instructions below carefully before starting on the questions

- All questions are about the past 2 weeks.
- For each question, please circle one number that best describes your situation.
- In case you never perform an activity:
  - Choose "no longer able (to)" if you no longer perform the activity because of limitations in the use of your arm.
  - When you are asked about an activity you never perform (or performed), please try to imagine whether you are limited in your ability to perform the activity.
- Some questions are about activities that you can perform with one hand. When answering these questions, please choose the arm with which you.
- Always performed this activity (before you had any complaints).

If you use aids or adapted equipment to perform an activity, please try to imagine how you would do without these aids.

Nr.	During the past 2 weeks, to what extent has MS …	Not at all	A little	Moderately	Quite a bit	Extremely	No longer able to
1	Limited your ability to write down a short sentence with a pen?	1	2	3	4	5	6
2	Limited your ability to grasp small objects, for example a key or a ballpoint pen?	1	2	3	4	5	6
3	Limited your ability to put on a coat?	1	2	3	4	5	6
4	Limited your ability to tie shoelaces?	1	2	3	4	5	6
5	Limited your ability to hold a full plate?	1	2	3	4	5	6
6	Limited your ability to pour from a bottle into a glass?	1	2	3	4	5	6
7	Limited your ability to turn the pages of a book?	1	2	3	4	5	6
8	Limited your ability to use a mouse of a computer?	1	2	3	4	5	6
9	Limited your ability to use a pen or pencil?	1	2	3	4	5	6
10	Limited your ability to turn a key in a lock?	1	2	3	4	5	6
11	Limited your ability to cut off a piece of paper with a pair of scissors?	1	2	3	4	5	6
12	Limited your ability to fasten a seatbelt in a car?	1	2	3	4	5	6
13	Limited your ability to fasten buttons?	1	2	3	4	5	6
14	Limited your ability to unbutton your shirt?	1	2	3	4	5	6
15	Limited your ability to take off a sweater or T-shirt?	1	2	3	4	5	6
16	Limited your ability to pick up coins from the table?	1	2	3	4	5	6
17	Limited your ability to use a keyboard?	1	2	3	4	5	6
18	Limited your ability to zip up a coat?	1	2	3	4	5	6
19	Limited your ability to carry a shopping bag?	1	2	3	4	5	6
20	Limited your ability to wash your hands?	1	2	3	4	5	6
21	Limited your ability to cut something with a knife?	1	2	3	4	5	6
22	Limited your ability to pierce food with a fork?	1	2	3	4	5	6
23	Limited your ability to dry off your body?	1	2	3	4	5	6
24	Limited your ability to open a bottle that has a screw cap?	1	2	3	4	5	6
25	Limited your ability to open a bottle of soft drink?	1	2	3	4	5	6
26	Limited your ability to wash the back of your shoulder?	1	2	3	4	5	6
27	Limited your ability to wash your hair?	1	2	3	4	5	6
28	Limited your ability to open a bag of crisps?	1	2	3	4	5	6
29	Limited your ability to bring a full glass or cup to your mouth?	1	2	3	4	5	6
30	Limited your ability to put toothpaste on a toothbrush?	1	2	3	4	5	6
31	Limited your ability to tuck a T-shirt/shirt in the back of your trousers using your hand?	1	2	3	4	5	6