Ashworth Scales are unreliable for the assessment of muscle spasticity

Noureddin Nakhostin Ansari\textsuperscript{a}; Soofia Naghdi\textsuperscript{b}; Hoda Moammeri\textsuperscript{c}; Shohreh Jalaie\textsuperscript{d}

\textsuperscript{a} Assistant Professor, Tehran University of Medical Sciences, Rehabilitation Faculty. Iran
\textsuperscript{b} Lecturer, currently PhD student in physiotherapy, Tehran University of Medical Sciences, Rehabilitation Faculty. Iran
\textsuperscript{c} Tehran University of Medical Sciences, Rehabilitation Faculty. Iran
\textsuperscript{d} Lecturer of Biostatistics, Tehran University of Medical Sciences, Rehabilitation Faculty, currently PhD student in Biostatistics. Iran

To cite this Article: Noureddin Nakhostin Ansari, Soofia Naghdi, Hoda Moammeri and Shohreh Jalaie, 'Ashworth Scales are unreliable for the assessment of muscle spasticity, Physiotherapy Theory and Practice, 22:3, 119 - 125
To link to this article: DOI: 10.1080/09593980600724188
URL: http://dx.doi.org/10.1080/09593980600724188
Ashworth Scales are unreliable for the assessment of muscle spasticity

Noureddin Nakhostin Ansari, PhD, PT, 1 Soofia Naghdi, MSc, PT, 2 Hoda Moammeri, BSc, PT, 3 and Shohreh Jalaie, MSc 4

1 Assistant Professor, Tehran University of Medical Sciences, Rehabilitation Faculty, Iran
2 Lecturer, currently PhD student in physiotherapy, Tehran University of Medical Sciences, Rehabilitation Faculty, Iran
3 Tehran University of Medical Sciences, Rehabilitation Faculty, Iran
4 Lecturer of Biostatistics, Tehran University of Medical Sciences, Rehabilitation Faculty, currently PhD student in Biostatistics, Iran

Ashworth Scales are the most widely used tests to assess the severity of muscle spasticity. These scales offer qualitative and subjective information; consequently, there are issues concerning validity and reliability. This article presents the results of a study comparing interrater reliability of the original and of the modified Ashworth Scales for measuring muscle spasticity in elbow flexors. Fifteen patients with hemiplegia (nine men and six women) with a median age of 52 years (interquartile range, 28–64) participated in this study. Two physiotherapists rated the muscle tone of elbow flexors according to ratings criteria of the Ashworth and the modified Ashworth Scales. Kappa values for the original Ashworth and the modified Ashworth Scales were 0.17 (SE = 0.21; p = 0.41) and 0.21 (SE = 0.12; p = 0.08), respectively. The scales showed similar levels of reliability ($\chi^2 = 0.0285$, df = 1, p = 0.7). The Ashworth Scales are not reliable for the assessment of muscle spasticity. In the absence of interrater reliability, the validity of the measurements may also be questioned.

Introduction

Spasticity is a common disabling feature of the Upper Motor Neuron (UMN) syndrome (Sehgal and McGuire, 1998; Lehmann et al, 1989). It is a problem that creates great difficulty for both patients and clinicians (Haley and Inacio, 1990; Katz and Rymer, 1989). Although spasticity may increase disability and is an impediment to the rehabilitation process (Barnes, 2003), weakness may present a more serious problem to movement dysfunction (Patten, Lexell, and Brown, 2004).

Spasticity may develop following an UMN lesion (e.g., brain injury, stroke, spinal cord injury, cerebral palsy, spinal degenerative diseases, etc.) (Sehgal and McGuire, 1998; Meythaler, 2001). From a clinical perspective, the term spasticity is difficult to define comprehensively (Young, 1994) and has often been used imprecisely (Bodine-Fowler and Botte, 1992). A widely accepted definition of spasticity states that “it is a velocity dependent increase in muscle tone with exaggerated tendon reflexes” (Lance, 1980). However, the clinical definition of spasticity and pathophysiology underlying spasticity remain unclear (Haley and Inacio, 1990; Sehgal and McGuire, 1998). The quantification of spasticity is as difficult as understanding its mechanism (Sehgal and McGuire, 1998; Sieb,
Clinicians perceive spasticity as a major impairment for individuals with UMN disorders (Bobath, 1990; Davis, 1985; Johnstone, 1983). Consequently, many therapists are interested in a valid and reliable tool (Pandyan et al, 1999; Pandyan, Price, Barnes, and Johnson, 2003; Katz and Rymer, 1989) to measure this phenomenon (Blackburn, Vliet, and Mockett, 2002).

Clinical tests are commonly used to quantify spasticity, although factors such as dystonia (Meythaler, 2001) and mechanical factors may also be involved (Dietz, Quintern, and Berger, 1981; O’Dwyer, Ada, and Neilson, 1996). The Ashworth Scale (1964) has been widely used clinically (Sehgal and McGuire, 1998; Johnson, 2001) to measure spasticity; however, it is a scale to document resistance to passive movement (RTPM) (Pandyan et al, 1999; Johnson, 2001). The Ashworth Scale provides an ordinal level measure of RTPM (Pandyan et al, 1999). Bohannon and Smith (1987) modified this scale and added the grade “1+.” The MAS is a nominal level measure of RTPM (Pandyan et al, 1999, 2001).

Despite efforts made, the validity and the reliability of the Ashworth (AS) and the modified Ashworth Scales (MAS) require further examination (Pandyan et al, 1999). Although these clinical scales are in general use (Ward, 2000), studies on their reliability produced conflicting results (Pandyan et al, 1999). There are also relatively few data available on the reliability of the Ashworth scales (Johnson, 2001). The previous studies, investigating reliability, have either used inappropriate methodology (Lee, Carson, Kinnine, and Patterson, 1989; Nuyens et al, 1994) or inappropriate statistics (Lee, Carson, Kinnine, and Patterson, 1989; Sloan et al, 1992; Bohannon and Smith, 1987; Nuyens et al, 1994). Hass, Bergstrom, Jamous, and Bennie (1996) compared the two scales in the lower limbs of 30 spinal cord injured subjects and found poor interrater reliability. Although the reliability was better for the Ashworth Scale ($\kappa = 0.41$ and $\kappa = 0.34$ for the AS and the MAS, respectively), the difference between the two was not significant. In the lower limbs of patients with stroke, measurements obtained with the MAS have been shown to have poor interrater reliability (Kendall tau-b = 0.62, $p = 0.461$) (Blackburn, Vliet, and Mockett, 2002). Brashear et al (2002), using the Ashworth Scale and the Disability Assessment Scale in the upper limbs of stroke subjects, found good reliability ($Kendall W = 0.598–0.792$).

Pandyan et al (1999) suggested that the testing procedures used in clinical trials needed to be described, and standardized procedures had to be used to improve the reliability. No study has so far compared the reliability between the two scales in hemiplegic patients. Therefore, the purpose of this study was to compare the reliability between the AS and MAS, by using a standard procedure, in assessing spasticity of elbow flexors in patients with hemiplegia.

Materials and methods

Subjects

Fifteen subjects with hemiplegia (nine men and six women), who could follow study instructions, were recruited for this study. The exclusion criteria were loss of range of motion in the elbow, contraindications to passive movements, nonconsent and taking tone-modifying drugs.

There were no age restrictions; two patients under 18 years of age were included in the study. The study was approved by the Research Council of Rehabilitation Faculty, Tehran University of Medical Sciences.

Procedure

Two physiotherapists participated in this study and provided all AS and MAS scores for all subjects. One physiotherapist had 10 years of clinical experience. The other had 1 year experience in practice. Neither had used the AS and MAS routinely before conducting this study, and no formal extensive training in the use of these scales had been offered to them. However, they had the descriptions of the ratings criteria. This situation was thought to best reflect the clinical situation because typical clinical procedures are normally based on written guidance and was unlikely to involve extensive training (Blackburn, Vliet, and Mockett, 2002).

Each patient was assessed during a single session. Patients were interviewed to determine age, gender, etiology, and duration of deficits.
Standardized written guidelines were as follows:

**Test area**
The testing area was quiet and screened off from other patients and therapists.

**Time of testing**
Testing took place between 10 AM and 11 AM.

**Starting position**
Patients were supine on a bed with head in midline and the arms alongside the trunk. Shoes were removed.

**Testing procedure**
Testing commenced 5 minutes after the subjects had been positioned. Patients were instructed to remain calm and to relax the arm completely. The forearm was grasped distally, just proximal to the wrist. The arm was stabilized proximal to the elbow. The forearm was in neutral supination. The patient’s elbow was extended from a position of maximal possible flexion to maximal possible extension by counting “one thousand and one” (Bohannon and Smith, 1987).

The **method of scoring**
The testing movement was repeated three times, and the raters were required to rate spasticity, by using both scales, based on first stretch.

The second assessor repeated the same procedure after a period of 5 minutes. The order of assessment between the two assessors was randomized, and no discussion of the results between the raters occurred during the study to ensure they were blind to each others results.

**Statistical analysis**
The level of agreement between the assessors was analyzed by using the Cohen’s Kappa test (Armitage, Berry, and Matthews, 2002). The difference in reliability between the two scales was analyzed with Chi-Square test (Fleiss, 1981). The kappa values were interpreted as suggested by Brennan and Silman (1992). The type 1 error rate was set to 0.05.

**Results**
The participants had hemiplegia due to cerebrovascular accident (11 subjects), tumor (3 subjects), or head trauma (1 subject). Eight patients had right hemiplegia, and seven subjects had left hemiplegia. The median age was 52 years (interquartile range, 28–64). The mean time since injury was 17.86 months (SD = 33.65).

Tables 1 and 2 illustrate the agreement between the raters for the AS and the MAS. For the AS, two raters agreed on eight patients (53.3%), and most agreement occurred for scores of 2 (33.3%) and 1 (20%), respectively. Most disagreement occurred between the grades 1 and 2 (33.3%). For the MAS, two raters agreed on seven patients (46.7%), and most agreement occurred for scores of 1+ (26.7%) and 1 (13.3%), respectively. Most disagreement occurred between the 1 and 1+ (20%) and the 1+ and 2 (20%).

Kappa values for both scales were poor, but kappa value for the MAS (κ = 0.21, SE = 0.12, p = 0.08) was slightly better than the AS (κ = 0.17, SE = 0.21, p = 0.41). There was no

| Table 1. Agreement between two raters using the AS for grading spasticity (n = 15). |
|---|---|---|---|---|
| Rater 1 | 0 | 1 | 2 | 3 | 4 |
| 0 | 1 |
| 1 | 3 | 2 |
| 3 | 3 | 5 |
| 4 | 1 |

κ = 0.17, SE = 0.12, p = 0.41.

| Table 2. Agreement between two raters using the MAS for grading spasticity (n = 15). |
|---|---|---|---|---|
| Rater 1 | 0 | 1+ | 2 | 3 | 4 |
| 0 | 1 |
| 1+ | 2 | 4 | 1 |
| 2 | 2 | 1 |
| 3 | 1 |
| 4 | |

κ = 0.21, SE = 0.12, p = 0.08.
significant difference between the reliability of the AS and MAS ($\chi^2 = 0.0285$, $df = 1$, $p = 0.7$).

**Discussion**

Previous studies to determine interrater reliability have used sample sizes ranging from 12 (Lee, Carson, Kinnine, and Patterson, 1989) to 36 (Blackburn, Vliet, and Mockett, 2002). The design of this study and the availability of patients willing or able to participate in this study limited the numbers possible for reliability assessment. A sample size of at least 12 patients is considered the minimum requirement to study reliability (Lee, Carson, Kinnine, and Patterson, 1989). Therefore, 15 subjects with hemiplegia were recruited for this study.

The results indicated that interrater reliability of both ASs (Ashworth, 1964; Bohannon and Smith, 1987) was only poor. The data suggest that agreement is only just better than chance further confirming the limited reliability of the ASs even when measuring spasticity in the upper limb. Despite the fact that the percentage agreement was higher for the AS, the kappa value was higher for the MAS. The kappa is the difference between observed and expected agreement. If there were no real agreement between the raters, then the expected frequencies would be determined by chance. In kappa calculation, agreement that could occur from chance have been taken into account (Armitage, Berry, and Mattews, 2002). The higher reliability of the MAS may be explained by the fact that this scale has more options to choose from and a scoring system with more grades would be best suited to detect differences (Ashby, Mailis, and Hunter, 1987).

Ample training in the use of the scale, interactions between assessors, and raters’ experience may have contributed to the high agreement between Bohannon and Smith (Johnson, 2001; Pandyan et al., 1999; Blackburn, Vliet, and Mockett, 2002; Hass, Bergstrom, Jamous, and Bennie, 1996). Brashear et al. (2002), Skold et al. (1998), and Bodin and Morris (1991) concluded that the scales would be reliable when used by trained and experienced raters. However, Pandyan et al. (1999) based on the raw data values provided by Bohannon and Smith (1987) and Bodin and Morris (1991) also calculated Cohen’s kappa values for these studies and concluded that MAS was moderately reliable at the elbow and the wrist flexors. In the present study, no formal training was offered to the raters because we wanted the procedure used to be a simulation of how the scale would be used in a clinical situation (Hass, Bergstrom, Jamous, and Bennie, 1996; Blackburn, Vliet, and Mockett, 2002). The results of the previous studies (Bohannon and Smith, 1987; Sloan et al., 1992; Lee, Carson, Kinnine, and Patterson, 1989; Brashear et al., 2002) suggest that reliability would greatly improve by training the scale users (Blackburn, Vliet, and Mockett, 2002). However, Pandyan et al. (2001), Pandyan, Price, Barnes, and Johnson (2003), and Bakheit et al. (2003) questioned the validity of the AS as a measure of spasticity.

The variation in raters’ experience may have had a role in poor reliability. The experienced assessor measured spasticity distributed across the entire scale but inexperienced assessor did not show this distribution and differentiation of scores. This is in agreement with Allison, Abraham, and Petersen (1996) who found their experienced raters measured spasticity across the scale. However, despite practice and interactions between the two experienced testers, Allison, Abraham, and Petersen (1996) demonstrated marginal reliability. In the present study, the less experienced rater reported that with increasing experience and better perception of spasticity, she could differentiate the scores. Skold et al. (1998) suggested that the experienced person should perform the Ashworth gradings. However, Blackburn, Vliet, and Mockett (2002) did not show good reliability despite the fact that their testers had more than 10 years of clinical experience. Although experience may improve the level of agreement, training may be a more important factor than clinical experience in reliability of the ASs.

In previous studies, repeated cycles of passive stretching prior to grading spasticity were used (Bohannon and Smith, 1987; Lee, Carson, Kinnine, and Patterson, 1989; Allison, Abraham, and Petersen, 1996). Repeated stretching causes variability in tone (Gregson et al., 1999). The raters in the present study commented that variability in tone affected their judgment in rating spasticity. Pandyan et al. (1999) recommended that repeated movements should be kept to a minimum. In this study, the movement was
repeated three times because one of the raters was inexperienced; we felt that once might not be sufficient for her to attribute a score (Blackburn, Vliet, and Mockett, 2002). In previous studies, they did not report how the raters scored the spasticity (Hass, Bergstrom, Jamous, and Bennie, 1996; Blackburn, Vliet, and Mockett, 2002; Bohannon and Smith, 1987; Allison, Abraham, and Petersen, 1996). Lee, Carson, Kinnine, and Patterson (1989) used the smallest of the five readings. Gregson et al (1999) took an average based on the four passive movements. Summation used by Lee, Carson, Kinnine, and Patterson (1989) and Nuyens et al (1994) is methodologically flawed (Pandyan et al, 1999). Averaging will also mask any possible unreliability. We believe for grading spasticity, clinicians should use only one passive stretching. We felt that one of the main important factors for poor reliability could be repeated passive stretching. Spasticity tends to change with passive stretching, and this factor makes the grading difficult for testers (Blackburn, Vliet, and Mockett, 2002). Skold et al (1998) used one passive stretching to score spasticity and found it correlated with electromyogram parameters. Pandyan et al (1999) described that the viscoelastic contributions to the RTPM would decrease with repeated cycles of stretching. Vattanasilp, Ada, and Crosbie (2000) reported that even one stretch decreased spasticity. Pandyan et al (2001) showed that RTPM tended to decrease with repeated measures, and mean RTPM of the first measure was higher than the second and third. Schmit, Dewald, and Rymer (2000) confirmed stretch reflex adaptation with repeated joint motion. Therefore, our raters were allowed to score spasticity based on the first stretching. However, performing three repetitions had a role in their confusion when grading spasticity. This method of scoring might have been another factor in this low reliability. Blackburn, Vliet, and Mockett (2002) reported that testing muscle three times was not always enough to establish the appropriate grade. Our raters commented that three times was sufficient to attribute a score and the variability due to repeated stretching made rating difficult.

With repeated testing, statistical regression also occurs, and all scores cluster toward the mean (Agostinucci, 1997). Statistical regression has been identified as a major source of secondary variation that may produce internal invalidity in an evaluation (Ottenbacher, 1986 cited by Agostinucci, 1997). Therefore, the evaluation’s validity decreases as the number of times the test increases (Agostinucci, 1997). We would recommend investigating the relationship between repeated measures with reliability of the evaluation and whether grading spasticity based on the first measure can improve interrater reliability.

We recognized that the method of analyzing might have been a factor in conflicting results. In different studies, methods of analysis were different (Bohannon and Smith, 1987; Gregson et al, 2000; Lee, Carson, Kinnine, and Patterson, 1989; Blackburn, Vliet, and Mockett, 2002; Brashear et al, 2002). Blackburn, Vliet, and Mockett (2002), assuming the MAS as an ordinal level of measurement, used Kendall tau-b for analysis. Gregson et al (2000) used weighted kappa test and concluded that the MAS is a reliable scale except at the ankle. However, Pandyan et al (1999) explained the MAS as a nominal level of measurement, and the weighted kappa is used where the categories are ordered (Armitage, Berry, and Matthews, 2002). In the study by Gregson et al, the kappa values were also reported. Based on these values, the reliability was poor to moderate. We used Chen’s kappa as suggested by Pandyan et al (1999), which computes perfect agreement. We recommend using kappa test for future studies. Cohen’s kappa statistic is probably the most appropriate measure of interrater reliability between two assessors using a scale with poorly defined ordinality (Pandyan et al, 1999) where there is no ordinal relationship between grades 1 and 1+ (Pandyan et al, 2001), and there are no significant differences between MAS 1, 1+, and 2 (Pandyan, Price, Barnes, and Johnson, 2003).

The subjects were graded 1 and higher more often than especially zero grade. Most agreement occurred on score 2 of the AS (33.3%) and score 1+ of the MAS (26.7%). This finding was not in agreement with others (Bohannon and Smith, 1987; Hass, Bergstrom, Jamous, and Bennie, 1996; Bodin and Morris, 1991; Blackburn, Vliet, and Mockett, 2002). The most agreement with higher grades points to this fact that the reliability may be greater in a more spastic patient group. However, this possible tendency requires further study with a more appropriate sample size. For the AS, most
agreement occurred between grades 1 and 2. For the MAS, disagreement occurred between 1 and 1+ and the 1+ and 2. This indicates that the reduction in the reliability of the ASs may be centered around the disagreement between grades 1 and 2. The poor agreement on MAS scores 1+ and 2 probably arises from their definition where the rater should discern the resistant point in the range of motion and perceive the “slightly increase” and “more marked increase” in tone at the same time. Therefore, adding an extra level of classification (1+) has increased the probability of errors occurring in the MAS (Pandyan et al, 1999). In addition, for the AS grades 1 and 2, the terms “slight increase,” “more marked increase,” and “part easily moved” makes discerning difficult. Terms used in the scales, especially in the MAS, are not clearly defined and are rich in ambiguity (Allison, Abraham, and Petersen, 1996). Therefore, the terminology used in the scales may have contributed to the poor reliability. To distinguish the grades in the scale from one another, we modified the scales as follows. A study to investigate the reliability of this modified form is underway.

0 No increase in tone
1 Slight increase in muscle tone, manifested by a catch and release or by minimal resistance at the end range of motion when the limb is moved in flexion or extension
2 Marked increase in tone, manifested by a catch in the middle range and resistance throughout the remainder of the range of motion, but limb easily moved
3 Considerable increase in tone, passive movement difficult
4 Limb rigid in flexion or extension

Conclusion
The study showed that when the AS and MAS were used in a standardized procedure and in a clinical situation in which the raters were not trained, the reliability of the both AS and MAS was poor. There was no significant difference between the two scales. Training may improve the agreement between testers, but the evidence questions the validity of the scales. When using Ashworth scales in the clinical setting to measure spasticity, one should consider the limitations of the scales.

References
Blackburn M, Vliet PV, Mockett SP 2002 Reliability of measurements obtained with the Modified Ashworth Scale in the lower extremities of people with stroke. Physical Therapy 82: 25–34

Davis PM 1985 Steps to follow. Berlin, Springer-Verlag


Pandyan AD, Price CIM, Barnes MP, Johnson GR 2003 A biomechanical investigation into the validity of the Modified Ashworth Scale as a measure of elbow spasticity. Clinical Rehabilitation 17: 290–294


Young RR 1994 Spasticity, A review. Neurology 44 (suppl 9): S12–S20