Online version at http://www.minervamedica.it

European Journal of Physical and Rehabilitation Medicine 2018 August;54(4):576-90
DOI: 10.23736/S1973-9087.17.04796-7

SYSTEMATIC REVIEW

Inter- and intra-rater reliability of the Modified Ashworth Scale: a systematic review and meta-analysis

Ana-Belén MESEGUER-HENAREJOS [1] *, Julio SÁNCHEZ-MECA [2], José-Antonio LÓPEZ-PINA [2], Ricardo CARLES-HERNÁNDEZ [1]

[1]Department of Physiotherapy, University of Murcia, Murcia, Spain; [2]Department of Basic Psychology and Methodology, University of Murcia, Murcia, Spain

*Corresponding author: Ana-Belén Meseguer-Henarejos, Department of Physiotherapy, University of Murcia, Murcia, Spain. E-mail: anabelen@um.es

ABSTRACT

INTRODUCTION: The Modified Ashworth Scale is the most widely clinical scale used to measure the increase of muscle tone. Reliability is not an immutable property of a scale and can vary as a function of the variability and composition of the sample to which it is administered. The best method to examine how the reliability of a test scores varies is by conducting a systematic review and meta-analysis of the reliability coefficients obtained in different applications of the test with the data at hand. The objectives of this systematic revision are: what is the mean inter- and intra-rater reliability of the Modified Ashworth Scale's scores in upper and lower extremities? Which study characteristics affect the reliability of the scores in this scale?

EVIDENCE ACQUISITION: The PubMed, Embase and CINAHL databases were searched from 1987 to February 2015. Two reviewers independently selected empirical studies published in English or in Spanish that applied the Modified Ashworth Scale and reported any reliability coefficient with the data at hand in children, adolescents or adults with spasticity.

EVIDENCE SYNTHESIS: Thirty-three studies reported any reliability estimate of Modified Ashworth Scale scores (N.=1065 participants). For lower extremities and inter-rater agreement, the mean intraclass correlation was ICC_+=0.686 (95% CI: 0.563 and 0.780) and for kappa coefficients, κ_+=0.360 (95% CI: 0.241 and 0.468); for intra-rater agreement: ICC_+=0.644 (95% CI: 0.543 and 0.726) and κ_+=0.488 (95% CI: 0.370 and 0.591). For upper extremities and inter-rater agreement: ICC_+=0.781 (95% CI: 0.679 and 0.853) and κ_+=0.625 (95% CI: 0.350 and 0.801); for intra-rater agreement: ICC_+=0.748 (95% CI: 0.671 and 0.809) and κ_+=0.593 (95% CI: 0.467 and 0.696). The type of design, the study focus, and the number of raters presented statistically significant relationships with ICC both for lower and upper extremities.

CONCLUSIONS: Inter- and intra-rater agreement for Modified Ashworth Scale scores was satisfactory. Modified Ashworth Scale' scores exhibited better reliability when measuring upper extremities than lower. Several characteristics of the studies were statistically associated to inter-rater reliability of the scores for lower and upper extremities.

(*Cite this article as:* Meseguer-Henarejos AB, Sánchez-Meca J, López-Pina JA, Carles-Hernández R. Inter- and intra-rater reliability of the Modified Ashworth Scale: a systematic review and meta-analysis. Eur J Phys Rehabil Med 2018;54:576-90. DOI: 10.23736/S1973-9087.17.04796-7)

KEY WORDS: Reproducibility of results - Meta-analysis - Muscle tonus - Weights and measures.

Introduction

The Modified Ashworth Scale is the most widely clinical scale used to measure the increase of muscle tone [1] which is manifested by an increased resistance of joints to passive movement.[2] The increase of muscle tone can be present in different pathologies such as stroke, multiple sclerosis, spinal cord injury, traumatic brain injury, cerebral palsy and in other neurological conditions which can cause upper motor neuron lesions.[3-7] If this muscle tone is not treated it can imply a shortening of muscles and connective tissue, and resulting in contractures and decreased range of active and passive joint motion. Further, it is associated with abnormal posture, weakness, pain, fatigability, pressure scores, sleep disturbances, decreased sense of security, limited mobility, self-care and domestic life, and diminished quality of life.[8-12]

This scale is applied manually to determine the per-

ceived resistance of muscles while moving a joint through its full range of movement. It is administered easily and in a short-time, and do not need equipment to rate muscle tone in clinical practice.[13-15]

The Modified Ashworth Scale is commonly used for obtaining baseline assessment of increased muscle tone, monitoring the course of disease, determining the effectiveness of pharmacologic and rehabilitation interventions to reduce overall muscle tone elevation, normalizing tone in selected muscle groups, and guiding physiotherapeutic and other treatment decisions.[13, 16-18]

From its original validation in the USA,[19] the Modified Ashworth Scale has been used to many cultures and countries, such as Canada,[20] UK,[21] Belgium,[22] The Netherlands,[23] Germany,[24] Switzerland,[25] Poland,[26] Turkey,[27] Finland,[28] Israel,[29] Iran,[30] Australia,[31] and China.[32, 33]

To be useful, a measurement tool must have good psychometric properties, such as reliability and validity. This study is focused on the reliability of the Modified Ashworth Scale scores. As the Modified Ashworth Scale is administered by a rater, its most important type of reliability is the inter-rater and the intra-rater agreement. Inter-rater reliability is obtained by applying the Modified Ashworth Scale to a sample of participants by two or more independent assessors. Intra-rater reliability is measured by having an assessor to measure the same participants on different occasions.

The reliability of the Modified Ashworth Scale was first tested in 30 adult patients with central nervous system lesions (closed head injuries, stroke, and multiple sclerosis) to calculate inter-rater reliability on the elbow flexor muscles.[19] Furthermore, the inter- and intra-rater reliability of the Modified Ashworth Scale has been studied in different muscles and different pathologies of the central nervous system, both in children and adult patients.[15, 25, 30, 34-36] These studies show a clear variation in both inter-rater and intra-rater reliability estimates.

Reliability is not an immutable property of a scale, but of the test scores applied to a given sample of participants and in a specific setting. Reliability of test scores can vary as a function of the variability and composition of the sample to which it is administered (*e.g*., clinical *vs*. nonclinical population, age, gender and ethnic distribution, language, test adaptation, etc.).[37] As reliability can vary from one test application to the next, the best method to examine how the reliability of a test scores varies is by conducting a systematic review and meta-analysis of the reliability coefficients obtained in different applications of the test with the data at hand. This kind of systematic review is usually named reliability generalization meta-analysis.[38]

The objectives of this systematic revision are:

- to find what is the mean inter- and intra-rater reliability of the Modified Ashworth Scale scores to measure the muscle tone of upper and lower extremities.
- to find which methodological and substantive characteristics of the studies can affect the reliability coefficients of the Modified Ashworth Scale scores.

Evidence acquisition

Study identification and selection

A literature search was undertaken to identify eligible studies for the systematic review. The PubMed, Embase and CINHAL databases were searched from 1987 (year of the publication of the scale) to February 2015. In these databases the sentence "Modified Ashworth Scale" was searched for in the title and abstract. No search terms were used for type of design or type of illness. The electronic search was complemented by checking the reference lists of included studies. The study selection process was conducted by two assessors who independently decided eligibility by titles and abstracts based on predetermined criteria. Where eligibility was unclear from the title and abstract, the full-text version was obtained and examined by both assessors. Disagreements were resolved by consensus, with a third assessor consulted when necessary.

To be included, the empirical studies (observational and experimental) had to apply the Modified Ashworth Scale to a sample of patients with spasticity, the patients had to pertain to a clinical population (children, adolescents or adults), and the study was required to be written in English or Spanish. Studies that applied any adaptation of the original Modified Ashworth Scale were included. Unpublished studies were also included. The studies had to report any inter- or intra-rater reliability coefficient with the data at hand.

Assessment of characteristics of selected studies

Methodological and substantive characteristics of the studies were extracted in order to examine the potential influence of moderator variables on the reliability estimates. The following methodological variables were coded: test version (original *vs*. adapted version), type of design (observational *vs*. experimental), study focus (psychometric

vs. applied), sample size, previous training of the raters with the Modified Ashworth Scale (yes *vs*. no), raters experience with the Modified Ashworth Scale (yes *vs*. no, and number of months), inter-rater interval (in days), number of raters, and number of replies. The following substantive variables were coded: age of the sample (coded as the mean in years and distinguishing between adults *vs*. children and adolescents), standard deviation of age (years), country and continent where the study was conducted, gender distribution (% male), target population (clinical *vs*. nonclinical), type of illness, extremities assessed (upper *vs*. lower), history of illness (mean and *SD* in years), training of the raters (physical therapist, physician, etc.), year of study, and publication source (published vs. unpublished).

Statistical analysis

To assess the reliability of the data extraction process, two raters independently coded the characteristics of all studies that fulfilled the selection criteria. For categorical moderator variables Cohen's kappas were calculated, whereas for the continuous ones intraclass correlations were computed. Regarding categorical variables, Cohen's kappas ranged from 0.41 to 1, with a mean of 0.84±0.19, whereas intraclass correlations calculated for continuous variables ranged from 0.68 to 1, with mean of 0.95±0.10. Inconsistencies between the raters were resolved by consensus.

Due to the specific nature of Modified Ashworth Scale, only agreement indices were included in the meta-analysis. In particular, the following types of reliability coefficients were extracted from the studies: intraclass, Pearson, Spearman, and Kendall's tau-b correlations, as well as weighted and unweighted Cohen's kappa coefficients, and squared weighted kappa. In addition, these reliability estimates were separately grouped depending on the inter-rater or intra-rater nature of the calculations. The different reliability coefficients were translated into Fisher's *Z* in order to normalize its distribution and stabilize the variances.

Separate meta-analyses were carried out for the different agreement coefficients. However, due to the wide variety of reliability coefficients found in the studies, intraclass, Pearson, Spearman, and squared weighted kappa coefficients were integrated into the same meta-analysis, as they can be considered as comparable coefficients.[39] Cohen's kappas, both weighted and unweighted, were included in the same meta-analysis. In addition, separate meta-analyses were carried out for the reliability coefficients obtained from upper and lower extremities, as the literature shows evidence that reliability can be different for different extremities. Thus, a total of six meta-analyses were planned for each type of extremity (upper and lower extremities): three meta-analyses for reliability estimates obtained from inter-rater agreement (intraclass correlations, Cohen's kappas, and Kendall's tau-b coefficients), and another three meta-analyses from intra-rater agreement.

Meta-analyses were done by assuming a random-effects model, so that each reliability estimate was weighted by its inverse variance, the variance of each coefficient being the sum of the within-study variance and the between-studies variance. Previous to meta-analytic integration, the individual reliability coefficients, *r*, were translated into the Fisher's *Z* by means of:

$$Z_r = 0.5\ Ln\ [(1 + r)/(1 - r)]$$

with *Ln* being the natural logarithm. The within-study variance of each coefficient transformed to Fisher's *Z* was obtained by:

$$V(Z_r) = 1 / (n - 3)$$

n being the sample size of the study. The between-studies variance was estimated by the DerSimonian and Laird's method. For each meta-analysis a weighted mean reliability coefficient was calculated, as well as a 95% confidence interval and heterogeneity statistics (Cochran's *Q* statistic and I^2 index). In order to facilitate the interpretation of the results, once statistical integration was carried out the Fisher's Zs were back transformed to the reliability coefficient metric, by means of:

$$r = (e^{2Z} - 1) / (e^{2Z} + 1)$$

with *e* being the base of the natural logarithms.[40] In order to determine whether publication bias might be a threat to the validity of the meta-analytic results, funnel plots with the Duval and Tweedie's[41] trim-and-fill imputation method were applied, as well as Egger tests.[42]

The meta-analysis with the largest number of reliability estimates was that of intraclass correlations obtained from inter-rater agreement, both for lower and upper extremities. Thus, analyses of moderator variables that can explain heterogeneity among the reliability estimates were conducted for these two datasets. For qualitative moderator variables, weighted ANOVAs were applied, with the Q_B statistic testing the significance of differences among the mean reliability coefficients of the different categories of the moderator. For continuous moderator variables,

simple meta-regressions were applied, with a Z statistic enabling us to test the significance of the moderator on the reliability estimates. In all cases mixed-effects models were assumed and the model misspecification was assessed with the Q_W and Q_E statistics for the ANOVAs and meta-regressions, respectively. The proportion of variance accounted for by each moderator variable was estimated by means of the R^2 index.[43-45] The statistical analyses were conducted with the Comprehensive Meta-analysis statistical software v. 3.0.[46]

Evidence synthesis

Figure 1 presents a flowchart describing the selection process of the studies that met selection criteria. Electronic databases consulted gave a total of 1066 references screened for their potential inclusion in the systematic review. In total, 642 references were empirical studies that applied the Modified Ashworth Scale. The remaining references were systematic reviews and meta-analyses, theoretical reviews of the literature, posters published in scientific journals, studies not written in English or Spanish, studies that did not apply the Modified Ashworth Scale, but did apply the Ashworth Scale or the Modified Ashworth Scale, study protocols, studies with animals, and a duplicated study.

Out of 636 references that applied the Modified Ashworth Scale only 33 studies (5.5%) reported some reliability estimate with the data at hand,[1, 4, 15, 19, 21-26, 29-33, 34-36, 47-60] the remaining 603 inducing it from previous references. These figures implied a dramatically large reliability induction rate for the Modified Ashworth Scale of 94.17%. The 33 studies that reported some reliability estimate constituted the database for this systematic review and meta-analysis.

Table I presents the characteristics of the studies. Some included inter- and intra-rater reliability,[1, 30] whereas others reported inter-rater reliability [19, 47] or intra-rater reliability only.[35, 48] With the exception of an N.=1 study,[49] the sample sizes of studies ranged from 6 [50, 51] to 93.[20]

The samples included patients with very different diagnoses. Some studies included patients with pathologies that course with spasticity, such as cerebral palsy,[29, 33] stroke,[52, 61] traumatic brain injury,[33, 34] spinal cord injury,[36, 53] or cerebral hypoxia.[54, 55] Others included patients with profound intellectual and multiple disabilities.[23] Most included patients with the same pathology,[24, 39] although some had samples of patients with different pathologies.[19, 30]

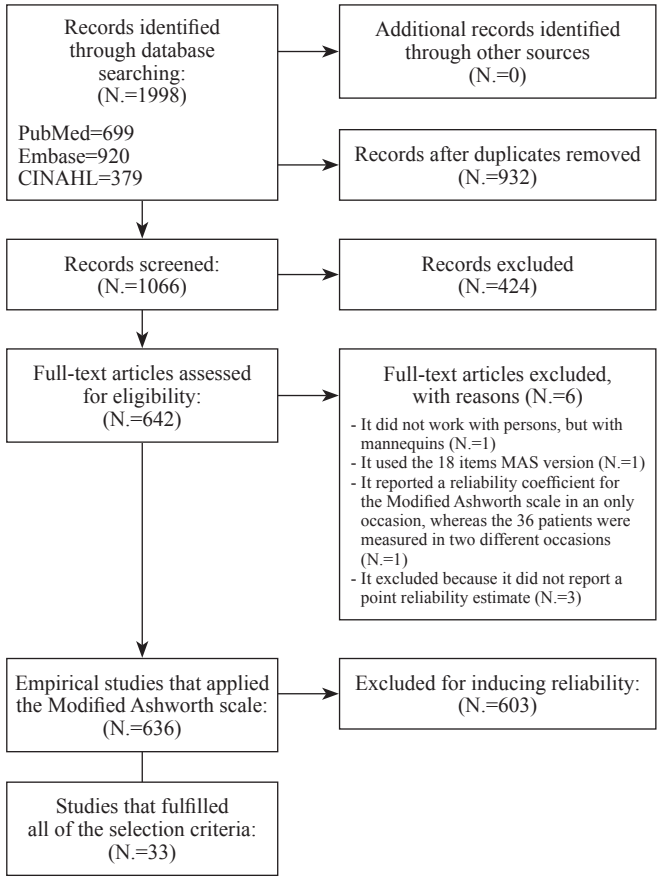

Figure 1.—Flow of studies through the systematic review.

Both upper and lower extremities and many different muscles have been assessed with the Modified Ashworth Scale. Some studies focused on only one muscle, such as the elbow flexors,[19, 52] knee extensors,[48] or ankle plantar flexors,[34, 56] whereas others examined several types of muscles in the upper,[22, 57] lower [36, 58] or in both extremities.[25, 58]

Mean reliability and heterogeneity

The 33 studies that reported at least a reliability estimate of the Modified Ashworth Scale gave 38 independent samples based on a total sample of 1065 participants (mean sample size=28±19.2). The studies reported different agreement reliability coefficients. Separate meta-analyses were conducted as a function of the extremities assessed (upper *vs*. lower), type of reliability (inter- *vs*. intra-rater), and type of reliability coefficient (intraclass *vs*. Cohen's

TABLE I.—*Summary of included studies.*

Study	Reliability examined	N.	Diagnosis	Setting	Age (years) Mean±SD	Joint/muscles	Time since injury (months) Mean±SD
Allison *et al.* (1996) [34]	Inter-rater Intra-rater	30	Traumatic brain injury	Healthcare Rehabilitation Center	28.3±10.8	Ankle plantar flexor	56.60±48.90
Ansari *et al.* (2008) [30]	Inter-rater Intra-rater	30	Traumatic brain injury Stroke	University	59.4±14.0	Shoulder adductors Elbow flexors Wrist flexors Hip adductors Knee extensors Ankle plantar flexors	NA
Bar-Haim *et al.* (2007) [29]	Inter-rater Intra-rater	78	Cerebral palsy	Outpatient rehabilitation clinics	9.5±3.8	Hip extensors Knee extensors	NA
Bohannon *et al.* (1987) [19]	Inter-rater	30	Traumatic brain injury Stroke Multiple sclerosis	Outpatient rehabilitation center	59.3±17.6	Elbow flexor	NA
Cheng *et al.* (2015) [35]	Intra-rater	10	Cerebral palsy	Local hospital Special education school	9.7±1.7	Knee extensors	NA
Clopton *et al.* (2005) [59]	Inter-rater Intra-rater	17	Cerebral palsy Developmental delay Traumatic brain injury	Outpatient rehabilitation center	7.0	Elbow flexors Hip adductors Knee quadriceps Knee gastrocnemius Ankle soleus	NA
Craven *et al.* (2010) [36]	Inter-rater Intra-rater	20	Spinal cord injury	Tertiary academic rehabilitation center	38.9±13.6	Hip adductors, abductors Knee quadriceps, hamstrings Ankle plantar flexors, dorsiflexors	106.7±96.0
Fasoli *et al.* (2008) [49]	Intra-rater	12	Cerebral palsy	Outpatient Rehabilitation Hospital	6.6	Shoulder adductors Elbow flexor, extensor, pronator, supinator Wrist flexors, extensors	6.0
Fasoli *et al.* (2008) [57]	Intra-rater	1	Cerebral palsy	Outpatient Rehabilitation Hospital		Shoulder adductors Elbow flexor, extensor, pronator, supinator Wrist flexors, extensors	NA
Fosang *et al.* (2003) [31]	Inter-rater Intra-rater	18	Cerebral palsy	Outpatient clinic Local special school	6.3	Hip adductors Knee hamstrings Ankle gastrocnemius	NA
Fragala *et al.* (2002) [58]	Inter-rater	7	Cerebral palsy	Outpatient rehabilitation hospital	6.7	Hip flexors, extensors, adductors Knee flexors, extensors Ankle plantar flexors	NA
Gregson *et al.* (1999) [52]	Inter-rater Intra-rater	32	Stroke	Acute stroke University hospital	74	Elbow flexors	1.6
Gregson *et al.* (2000) [1]	Inter-rater Intra-rater	35	Stroke	Acute stroke University hospital	73	Elbow flexors Wrist flexors Knee flexors Ankle plantar flexors	1.3
Haas *et al.* (1996) [53]	Inter-rater	30	Spinal cord injury	Outpatient rehabilitation center	40.3	Hip flexors, extensors, adductors Ankle plantar flexors	17.2
Hagenbach *et al.* (2007) [25]	Inter-rater	22	Spinal cord injury	Outpatient hospital	40.9	Shoulder adductors, abductors Elbow flexors, extensors Wrist flexors, extensors Hip flexors, extensors Knee flexors, extensors	13.3
Hesse *et al.* (2003) [24]	Inter-rater	12	Stroke	Community rehabilitation center	63.6	Elbow flexors Wrist flexors Fingers flexors	9.3

(*To be continued*)

TABLE I.—*Summary of included studies (continues).*

Study	Reliability examined	N.	Diagnosis	Setting	Age (years) Mean±SD	Joint/muscles	Time since injury (months) Mean±SD
Kaya *et al.* (2011) [47]	Inter-rater	64	Stroke	Outpatient rehabilitation clinic	60.5±11.9	Elbow flexors	3.9
Klingels *et al.* (2010) [22]	Inter-rater Intra-rater	30 23	Cerebral palsy	University Hospital	10.5±2.6 NA	Shoulder flexors, adductors, abductors Elbow flexors, extensors Wrist flexors, extensors Pronators, supinators	NA
Li *et al.* (2014) [61]	Inter-rater Intra-rater	51	Stroke	Inpatients rehabilitation hospital	59.0±14.6	Elbow flexors Ankle plantar flexors	3.7±4.3
Mehrholz *et al.* (2005) [54]	Inter-rater Intra-rater	50	Traumatic brain injury Stroke Cerebral hypoxia	Early rehabilitation center	58.2±14.1	Shoulder extensors, internal rotators Elbow flexors, extensors Wrist flexors, extensors Hip flexors, extensors Knee flexors, extensors Ankle plantar flexors	2.1±1.1
Mehrholz *et al.* (2005) [55]	Inter-rater Intra-rater	30	Traumatic brain injury Stroke Cerebral hypoxia	Rehabilitation center	63.9±12.9	Shoulder extensors Elbow flexors Wrist flexors Knee flexors Ankle plantar flexors	2.6±3.1
Motl *et al.* (2006) [56]	Inter-rater	27	Multiple sclerosis	University	44.9±8.3	Ankle plantar flexors	99.6±68.4
Mutlu *et al.* (2008) [15]	Inter-rater Intra-rater	30	Cerebral palsy	University	4.4±1.6	Hip flexors, adductors, internal rotators Knee flexors Ankle plantar flexors	NA
Numanoglu *et al.* (2012)[48]	Intra-rater	37	Cerebral palsy	University	9.0±4.4	Elbow flexors Wrist flexors Hip adductors Knee flexors Ankle plantar flexors	NA
Salem *et al.* (2010) [51]	Intra-rater	6	Cerebral palsy	University	6.5±2.5	Ankle plantar flexors	NA
Sloan *et al.* (1992) [21]	Inter-rater	34	Traumatic brain injury Stroke	Rehabilitation medicine unit, hospital	58.8±17.8	Elbow flexors, extensors Knee flexors, extensors	NA
Smith *et al.* (2002) [14]	Inter-rater	23	Spinal cord injury	Tertiary care outpatient and inpatient spinal cord injury rehabilitation center	33.4±12.5	Knee flexors, extensors	29.8±43.2
Tederko *et al.* (2007) [26]	Inter-rater Intra-rater	30	Spinal cord injury	Acute rehabilitation center	33.9	Shoulder, elbow, wrist, fingers, hip, knee, ankle (does not specify muscles)	4.0
Waninge *et al.* (2011) [23]	Inter-rater Intra-rater	23 35	Profound intellectual and multiple disabilities	NA	33.5 NA	Upper and low limbs (does not specify muscles)	NA
Yam *et al.* (2006) [32]	Inter-rater	17	Cerebral palsy	Department of Physiotherapy and Pediatrics and Adolescent Medicine in Hospital	7.8	Hip adductors Ankle plantar flexors	NA
Kelly *et al.* (2008) [60]	Inter-rater	10	Cerebral palsy	Outpatient clinic	4.1±1.2	Knee flexors Ankle plantar flexors	NA
Kim *et al.* (2011) [50]	Inter-rater	6	Cerebral palsy	Rehabilitation Medicine Department, Clinical Center	12.5±4.1	Elbow (NA)	NA
Allison *et al.* (1995) [33]	Intra-rater	34	Traumatic brain injury	Healthcare Rehabilitation Center	30.3±10.8	Ankle plantar flexors	45.6

NA: not available.

kappa). Thus, a total of eight meta-analyses were conducted. Kendall's tau-b coefficients were reported in only two and three studies for upper and lower extremities, respectively. The small number of this type of coefficients did not enable us to conduct meta-analyses.

Table II presents the results of each of the eight meta-analyses conducted. Regarding lower extremities, 13 studies reported some kind of intraclass correlation from inter-rater agreement, with a mean coefficient of ICC_+=0.686 (95% CI: 0.563 and 0.780) and a large heterogeneity (I^2=68.1%). Figure 2 presents a forest plot of these coefficients. The eight studies that reported some kappa coefficient for inter-rater reliability (Figure 3) exhibited a mean coefficient of κ_+=0.360 (95% CI: 0.241 and 0.468) and a null heterogeneity (I^2=0%). In 12 studies an intraclass correlation obtained from intra-rater agreement was reported (Figure 4). This meta-analysis presented a mean coefficient of ICC_+=0.644 (95% CI: 0.543 and 0.726), very similar to that obtained from inter-rater agreement. The heterogeneity exhibited by these coefficients was medium (I^2=42.5%). Finally, seven studies reported some kappa coefficient from intra-rater agreement (Figure 5). The mean kappa coefficient was κ_+=0.488 (95% CI: 0.370 and 0.591) and the coefficients exhibited a null heterogeneity (I^2=0%).

Regarding upper extremities, Table II shows that 11 studies reported some kind of intraclass correlation from inter-rater agreement. Figure 6 presents a forest plot with these intraclass correlations, that exhibited a mean coefficient of ICC_+=0.781 (95% CI: 0.679 and 0.853) and a large heterogeneity (I^2=69.8%). Six studies that reported some kappa coefficient for inter-rater agreement (Figure 7) presented a mean coefficient of κ_+=0.625 (95% CI: 0.350 and 0.801) and a large heterogeneity (I^2=83.4%). In 11 studies an intraclass correlation obtained from intra-rater agreement was reported (Figure 8). This meta-analysis ob-

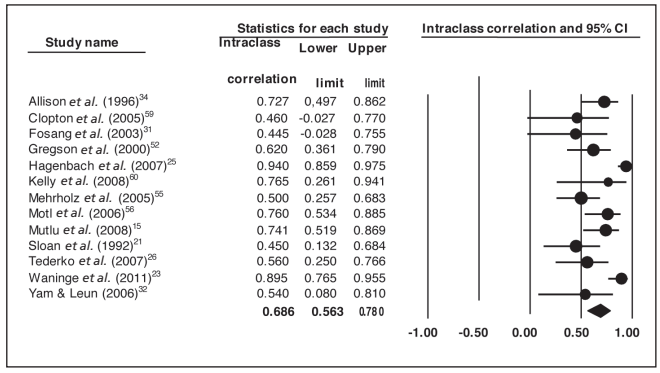

Figure 2.—Forest plot of the intraclass correlation coefficients obtained from inter-rater reliability estimates for lower extremities.

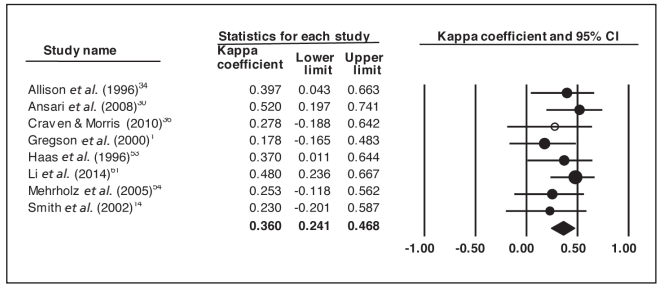

Figure 3.—Forest plot of the kappa coefficients obtained from inter-rater reliability estimates for lower extremities.

TABLE II.—*Synthesis of the reliability estimates obtained from the different inter- and intra-coder agreement methods.*

Type of reliability	N.	Min.	Max.	Mean	95% CI (range)	Q	P value	I^2
Lower extremities								
Inter-rater reliability:								
Intraclass correlation	13	0.445	0.940	0.686	0.56-0.78	37.62	<0.001	68.1
Cohen's kappa	8	0.178	0.520	0.360	0.24-0.47	4.32	0.742	0.0
Intra-rater reliability:								
Intraclass correlation	12	0.428	0.920	0.644	0.54-0.73	19.14	0.059	42.5
Cohen's kappa	7	0.330	0.594	0.488	0.37-0.59	1.83	0.934	0.0
Upper extremities								
Inter-rater reliability:								
Intraclass correlation	11	0.510	0.940	0.781	0.68-0.85	33.17	<0.001	69.8
Cohen's kappa	6	0.307	0.880	0.625	0.35-0.80	30.09	<0.001	83.4
Intra-rater reliability:								
Intraclass correlation	11	0.490	0.880	0.748	0.67-0.81	15.77	0.106	36.6
Cohen's kappa	4	0.438	0.690	0.593	0.47-0.70	3.24	0.357	7.3

N.: number of studies; Min.: minimum reliability coefficient; Max.: maximum reliability coefficient; Q: Cochran's statistic to test the null hypothesis of homogeneity; I^2: heterogeneity index.

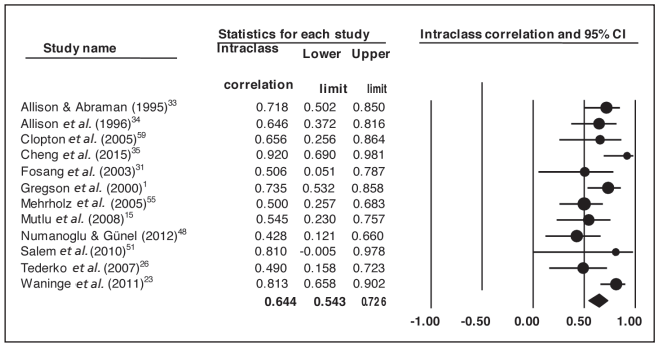

Figure 4.—Forest plot of the intraclass correlation coefficients obtained from intra-rater reliability estimates for lower extremities.

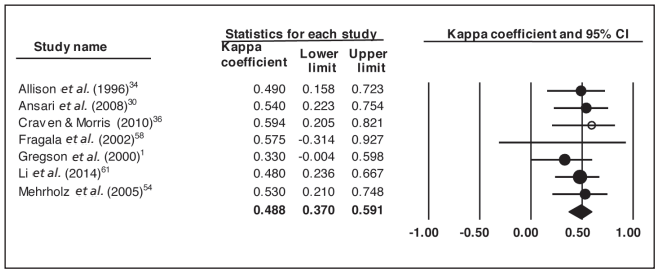

Figure 5.—Forest plot of the kappa coefficients obtained from intra-rater reliability estimates for lower extremities.

tained a mean coefficient of ICC_+=0.748 (95% CI: 0.671 and 0.809), like that obtained from inter-rater agreement. Heterogeneity exhibited by these coefficients was medium (I^2=36.6%). Finally, only four studies reported some kappa coefficient from intra-rater agreement (Figure 9), with a mean of κ_+=0.593 (95% CI: 0.467 and 0.696) and a low heterogeneity (I^2=7.3%).

Analysis of publication bias

In order to examine whether publication bias might confound the mean reliability coefficients obtained in the different meta-analyses, Egger tests and funnel plots with the trim-and-fill method were applied. Table III presents the results of applying the Egger test to each of the eight meta-analyses. The absence of statistical significance for the intercept in all is reason to discard publication bias. In addition, funnel plots were constructed and the trim and fill method for imputing missing values developed by Duval and Tweedie was applied.[41] Figure 10A-D, for lower extremities, and Figure 11A-D, for upper ones, present the funnel plots for each of the eight meta-analyses. Apart from two cases, the trim and fill method did not impute

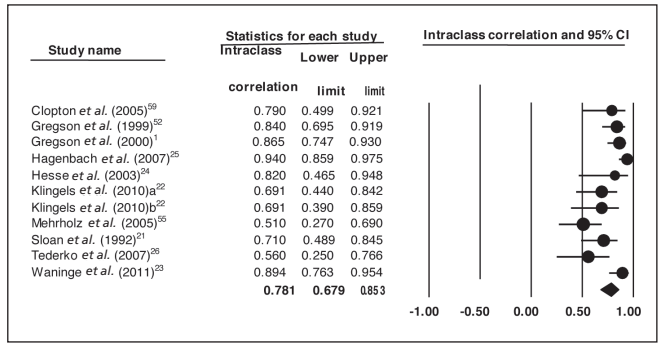

Figure 6.—Forest plot of the intraclass correlation coefficients obtained from inter-rater reliability estimates for upper extremities.

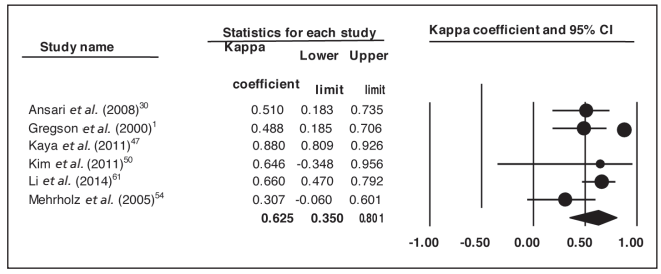

Figure 7.—Forest plot of the kappa coefficients obtained from inter-rater reliability estimates for upper extremities.

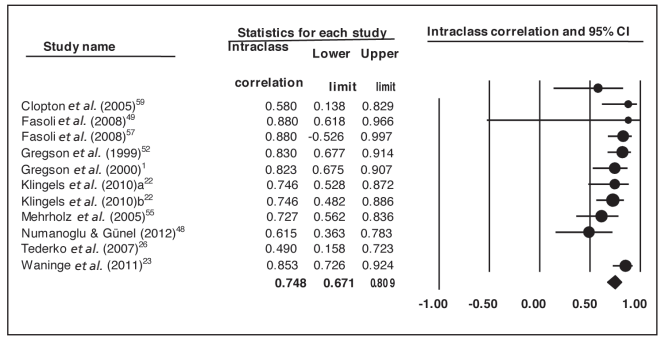

Figure 8.—Forest plot of the intraclass correlation coefficients obtained from intra-rater reliability estimates for upper extremities.

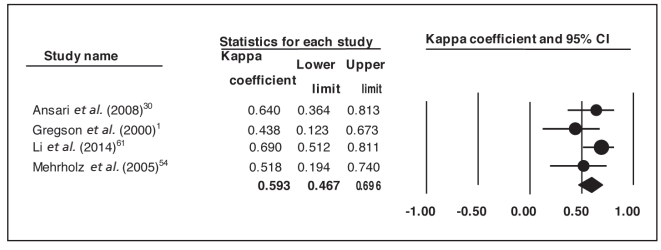

Figure 9.—Forest plot of the kappa coefficients obtained from intra-rater reliability estimates for upper extremities.

TABLE III.—*Results of the Egger tests applied to each of the eight meta-analyses to examine publication bias.*

Type of reliability	N.	b_0	SE	t	DF	P value
Lower extremities						
Inter-rater reliability:						
Intraclass correlation	13	2.040	2.310	0.88	11	0.396
Cohen's kappa	8	-2.343	1.802	-1.30	6	0.241
Intra-rater reliability:						
Intraclass correlation	12	1.729	1.297	1.33	10	0.212
Cohen's kappa	7	0.739	0.757	0.98	5	0.374
Upper extremities						
Inter-rater reliability:						
Intraclass correlation	11	4.080	2.575	1.58	9	0.147
Cohen's kappa	6	-2.973	3.034	-0.98	4	0.382
Intra-rater reliability:						
Intraclass correlation	11	0.480	1.274	0.38	9	0.715
Cohen's kappa	4	-4.507	4.101	-1.10	2	0.386

N.: number of studies: b_0: intercept of the unweighted simple regression of the standard error of reliability estimates on the reliability estimates; SE: standard error of b_0; *t*: statistic for testing the statistical significance of b_0; DF: degrees of freedom of the *t* statistic.

values to give symmetry to the funnel plot, meaning that publication bias can be discarded as a threat to the meta-analytic results. The only two cases in which the trim and fill method imputed three and two values, respectively, were in the meta-analyses for lower extremities with intra-rater intraclass correlations and Cohen's kappas. In both cases, however, the impact of adding missing values on the mean reliability was negligible. Therefore, publication bias can be discarded as a threat against the meta-analytic results.

Analysis of moderator variables

Of the four meta-analyses conducted for lower extremities, that with the largest number of studies and which exhibited heterogeneous reliability coefficients was intraclass correlations from inter-rater agreement. Thus, a search for moderator variables was carried out to explain at least part

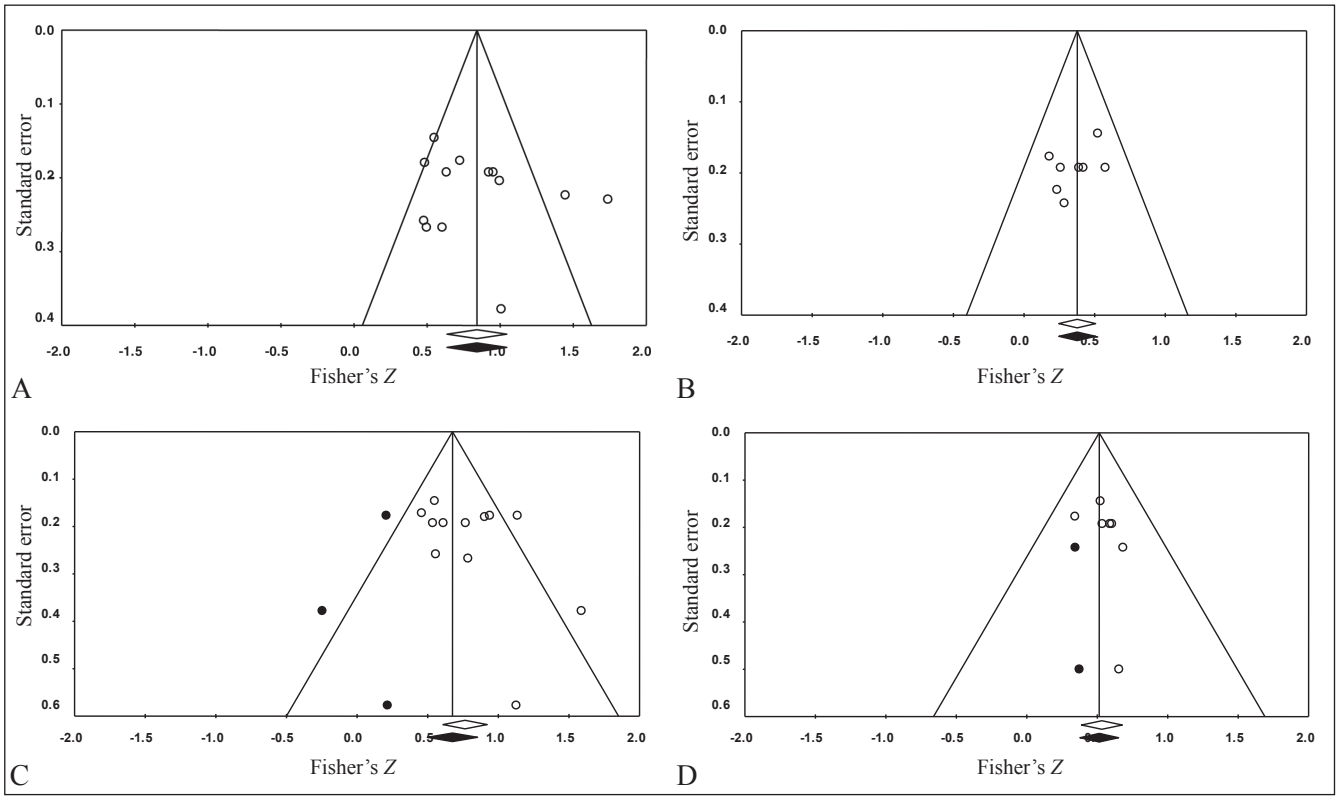

Figure 10.—A) Funnel plot with the trim-and-fill method of the intraclass correlation coefficients obtained from inter-rater reliability estimates for lower extremities. B) Funnel plot with the trim-and-fill method of the kappa coefficients obtained from inter-rater reliability estimates for lower extremities. C) Funnel plot with the trim-and-fill method of the intraclass correlation coefficients obtained from intra-rater reliability estimates for lower extremities. D) Funnel plot with the trim-and-fill method of the kappa coefficients obtained from intra-rater reliability estimates for lower extremities.

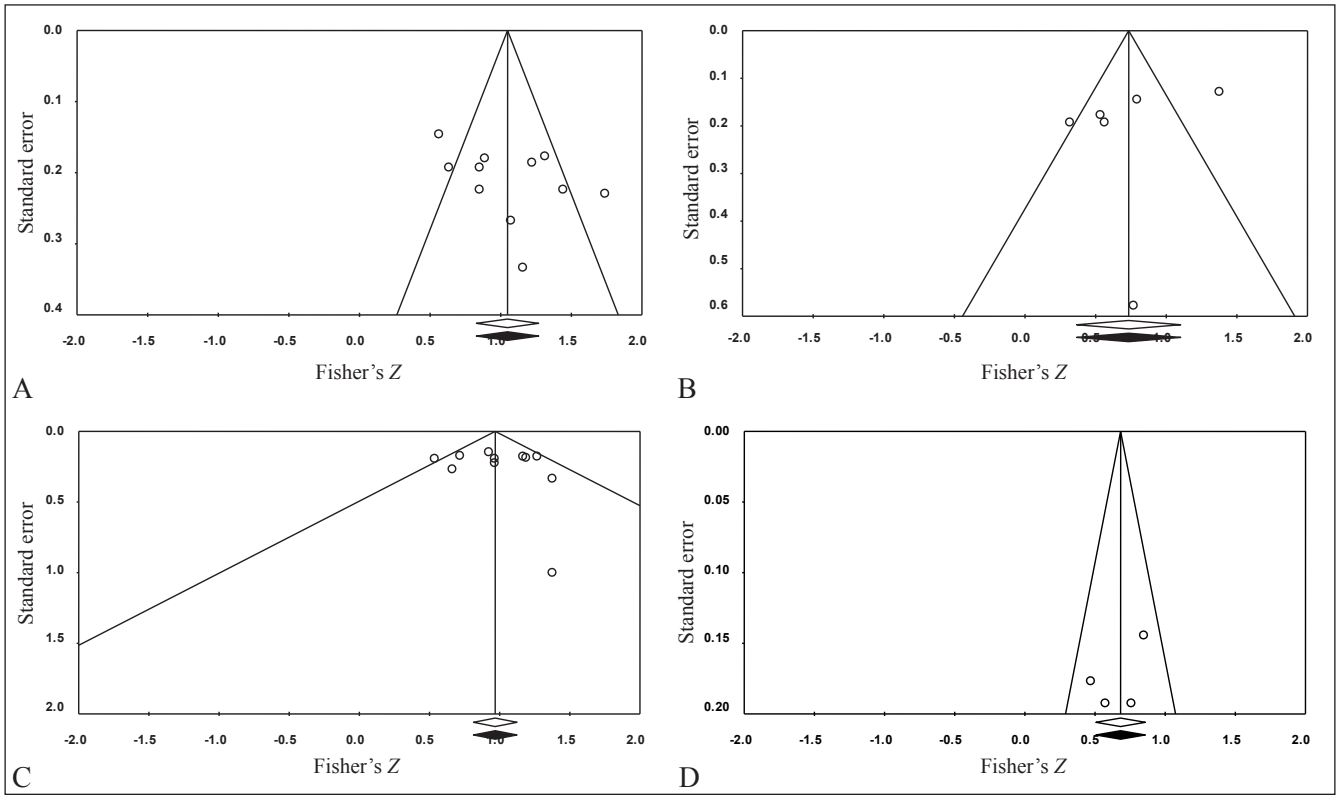

Figure 11.—A) Funnel plot with the trim-and-fill method of the intraclass correlation coefficients obtained from inter-rater reliability estimates for upper extremities. B) Funnel plot with the trim-and-fill method of the kappa coefficients obtained from inter-rater reliability estimates for upper extremities. C) Funnel plot with the trim-and-fill method of the intraclass correlation coefficients obtained from intra-rater reliability estimates for upper extremities. D) Funnel plot with the trim-and-fill method of the kappa coefficients obtained from intra-rater reliability estimates for upper extremities.

of the heterogeneity of the reliability estimates. Regarding upper extremities, an analysis of potential moderator variables was also done for intraclass correlations under inter-rater agreement.

Table IV presents the weighted ANOVAs conducted for the qualitative moderator variables on intraclass correlations obtained when using the MAS to assess lower extremities. The type of design exhibited a statistically significant relationship with intraclass correlations (P=0.002, R^2=0.61), the experimental studies showing a larger mean reliability (ICC_+=0.940) than the observational ones (ICC_+=0.644). However, this result must be interpreted cautiously, as only one study applied an experimental design, the remaining being observational. Classifying the studies as a function of its psychometric versus applied focus, also reached a statistically significant result (P=0.014, R^2=0.42), the applied studies showing a larger mean reliability (ICC_+=0.856) than that of psychometric (ICC_+=0.624). It is worth noting that the type of design and the study focus exhibited a large collinearity (φ=0.527). The remaining moderator variables presented in Table III did not reach a statistically significant relationship with intraclass correlations.

To examine the influence of continuous moderator variables, simple meta-regressions were applied. Table V presents the results of these analyses. Of the nine moderator variables analyzed, the number of raters reached a negative, statistically significant relationship with intraclass correlations (P=0.031, R^2=0.29), so that the larger the number of raters used in the study the lower the intraclass correlation. In addition, the standard deviation of the age presented a negative, statistically significant relationship with reliability (P=0.017, R^2=0.99); this result must be interpreted very cautiously due to the small number of studies on which it was based (6 only). None of the remaining variables reached statistical significance.

TABLE IV.—*Results of the ANOVAs for the qualitative methodological and substantive moderator variables on intraclass correlations obtained from inter-rater reliability in lower extremities.*

Moderator variable	N.	Mean	95% CI (range)	ANOVA results
Test version				$Q_B(1)=1.36$, P=0.243
Original version	7	0.619	0.41 to 0.77	$R^2=0.0$
Adapted version	6	0.749	0.58 to 0.85	$Q_W(11)=35.13$, P<0.001
Type of design				$Q_B(1)=9.91$, P=0.002
Observational	12	0.644	0.54 to 0.73	$R^2=0.61$
Experimental	1	0.940	0.82 to 0.98	$Q_W(11)=20.30$, P=0.041
Study focus				$Q_B(1)=6.04$, P=0.014
Psychometric	10	0.624	0.49 to 0.73	$R^2=0.42$
Application	3	0.856	0.71 to 0.93	$Q_W(11)=24.61$, P=0.010
Experience with MAS				$Q_B(2)=0.45$, P=0.797 $R^2=0.0$
Yes, all	2	0.636	0.26 to 0.84	$Q_W(2)=3.15$, P=0.207
Yes, someone	1	0.460	-0.11 to 0.80	
No	2	0.610	0.37 to 0.77	
Continent				$Q_B(3)=1.22$, P=0.749
Asia	1	0.540	-0.24 to 0.90	$R^2=0.0$
Europe	7	0.720	0.55 to 0.83	$Q_W(9)=35.01$, P<0.001
North America	4	0.694	0.41 to 0.85	
Oceania	1	0.445	-0.35 to 0.87	
Target population				$Q_B(2)=1.42$, P=0.492
Children and adolescents	5	0.605	0.31 to 0.79	$R^2=0.0$ $Q_W(10)=35.33$, P<0.001
Adults	7	0.741	0.58 to 0.84	
Both	1	0.560	-0.12 to 0.88	

N.: number of studies in each category; Q_B: between-category statistic to test the mean intraclass correlations; Q_W: total within-category statistic to test the model misspecification; R^2: proportion of variance explained by the moderator variable. It was not possible to analyze the experience with MAS because of not having sufficient data.

TABLE V.—*Results of the simple meta-regressions applied on the continuous moderator variables for intraclass correlations obtained from inter-rater reliability in lower extremities.*

Moderator variable	N.	b_j	Z	P value	Q_E	R^2
Methodological variables:						
Raters experience, months	6	0.001	1.17	0.243	2.48	0.0
Inter-rater interval, days	9	-0.025	-0.29	0.769	26.48**	0.0
Number of raters	13	-0.146	-2.16	0.031	27.60**	0.29
Number of replies	11	0.017	0.22	0.828	36.41**	0.0
Substantive variables:						
Mean age of the sample, years	11	0.0006	0.10	0.921	28.58**	0.0
SD of the age, years	6	-0.034	-2.40	0.017	2.17	0.99
Gender, % male	12	0.004	0.90	0.366	27.81**	0.0
Mean disorder history, years	6	0.002	0.46	0.642	19.60**	0.0
Publication year	13	0.031	1.67	0.094	30.59**	0.15

N.: number of studies; b_j: regression coefficient of the moderator variable; *Z*:= statistical test of b_j; Q_E: residual heterogeneity statistic to test the model misspecification; R^2: proportion of variance explained by the moderator variable. *P<0.05; **P<0.01.

Regarding upper extremities, Table VI presents the weighted ANOVAs conducted for the qualitative moderator variables on inter-rater intraclass correlations for the Modified Ashworth Scale. The type of design exhibited a marginally statistically significant relationship with intraclass correlations (P=0.060, R^2=0.31), with the experimental studies showing a larger mean reliability (ICC_+=0.906) than the observational ones (ICC_+=0.748). However, it is worth pointing out that the type of design was confounded with the study focus, as seen in Table VI. Therefore, these results must be interpreted cautiously. None of the remaining moderator variables included in Table V reached statistical significance.

The results of applying simple meta-regressions for continuous moderator variables are presented in Table VII. The number of raters reached a negative, statistically significant relationship with intraclass correlations (P=0.041, R^2=0.35), so that the larger the number of raters used in the study the lower the intraclass correlation. None of the remaining variables reached statistical significance.

TABLE VI.—*Results of the ANOVAs for the qualitative methodological and substantive moderator variables on intraclass correlations obtained from inter-rater reliability in upper extremities.*

Moderator variable	N.	Mean	95% CI (range)	ANOVA results
Test version				$Q_B(1)=0.28$, P=0.599
Original version	4	0.810	0.64 to 0.90	$R^2=0.0$
Adapted version	7	0.763	0.62 to 0.86	$Q_W(9)=30.30$, P<0.001
Type of design				$Q_B(1)=3.53$, P=0.060
Observational	9	0.748	0.64 to 0.83	$R^2=0.31$
Experimental	2	0.906	0.76 to 0.97	$Q_W(9)=23.78$, P=0.005
Study focus				$Q_B(1)=3.53$, P=0.060
Psychometric	9	0.748	0.64 to 0.83	$R^2=0.31$
Application	2	0.906	0.76 to 0.97	$Q_W(9)=23.78$, P=0.005
Previous training with MAS				$Q_B(1)=0.05$, P=0.828 $R^2=0.0$
Yes	4	0.718	0.56 to 0.83	$Q_W(4)=8.58$, P=0.072
No	2	0.691	0.42 to 0.85	
Experience with MAS				$Q_B(1)=0.13$, P=0.722 $R^2=0.0$
Yes, someone	1	0.790	-0.73 to 0.99	
No	1	0.510	-0.88 to 0.99	
Continent				$Q_B(2)=0.003$, P=0.957
Europe	10	0.781	0.67 to 0.86	$R^2=0.0$
North America	1	0.790	0.24 to 0.96	$Q_W(9)=33.10$, P<0.001
Target population				$Q_B(2)=2.29$, P=0.319
Children and adolescents	3	0.724	0.44 to 0.87	$R^2=0.0$ $Q_W(8)=27.45$, P=0.001
Adults	7	0.823	0.71 to 0.89	
Both	1	0.560	-0.09 to 0.88	

N.: number of studies in each category; Q_B: between-category statistic to test the mean intraclass correlations; Q_W: total within-category statistic to test the model misspecification; R^2: proportion of variance explained by the moderator variable.

TABLE VII.—*Results of the simple meta-regressions applied on the continuous moderator variables for intraclass correlations obtained from inter-rater reliability in upper extremities.*

Moderator variable	N.	b_j	Z	P value	Q_E	R^2
Methodological variables						
Inter-rater interval, days	7	0.542	1.95	0.051	14.33*	0.41
N. of raters	10	-0.164	-2.04	0.041	21.52**	0.35
N. of replies	9	0.029	0.31	0.759	27.10**	0.0
Substantive variables						
Mean age of the sample, years	8	0.001	0.19	0.849	24.51**	0.0
Gender, % male	9	0.008	0.86	0.389	26.36**	0.0
Mean disorder history, years	6	0.053	1.31	0.191	20.38**	0.14
Publication year	11	0.0009	0.04	0.967	33.13**	0.0

N.: number of studies; b_j: regression coefficient of the moderator variable; *Z*: statistical test of b_j; Q_E: residual heterogeneity statistic to test the model misspecification; R^2: proportion of variance explained by the moderator variable. It was not possible to analyze the raters experience nor the standard deviation of age because of not having sufficient data.
*P<0.05; **P<0.01.

Discussion

Reliability is not an invariant property of a measurement tool, but changes as a function of study characteristics, such as sample composition and variability.[62-64] There is an extended practice among researchers of inducing reliability of test scores from previous applications of the test (*e.g.*, from the original validation study). This erroneous practice, named "reliability induction," [65] does not consider that reliability can change from one test application to the next. Against this practice, it is advising researchers to report reliability estimates of test scores with the data at hand, although the focus of the study was not psychometric.[66]

This school of thought has favored carrying out numerous reliability generalization meta-analyses aimed at investigating how the reliability of test scores varies in different test applications and which factors can explain that variability. This investigation is the first systematic review and meta-analysis about the inter- and intra-rater reliability of the Modified Ashworth Scale scores. This research is important as clinicians and researchers need measurement tools capable of accurately assessing muscle tone in patients with upper motor neuron lesions and evaluating the success or failure of rehabilitative interventions.[34]

Our results point toward the existence of a satisfactory mean reliability of the Modified Ashworth Scale scores, both for inter- and intra-rater agreement. Following Brennan and Silman,[67] intraclass correlations under 0.5, between 0.5 and 0.75, and over 0.75, can be interpreted as reflecting a low, moderate, and high reliability, respectively. Thus, the mean intraclass correlations obtained for lower extremities in our meta-analysis exhibited a moderate magnitude (ICC_+=0.686 and 0.644 for inter- and intra-rater reliability, respectively), and a high magnitude for upper extremities (ICC_+=0.781 and 0.748, respectively). Regarding Cohen's kappa, Portney and Watkins proposed interpreting the clinical significance of kappa coefficient following this guide:[68] poor (<0.21), fair (0.21-0.40), moderate (0.41-0.60), good (0.61-0.80), and very good (0.81-1.00). Thus, in our meta-analysis mean kappas were of fair to moderate relevance for inter- and intra-rater reliability with lower extremities (κ_+=0.360 and 0.488, respectively), and good to moderate for upper extremities (κ_+=0.625 and 0.593, respectively). Our results in Table II also show evidence that the Modified Ashworth Scale scores exhibited better reliability for upper than for lower extremities. One possible explanation is that the lower limbs have greater length and overall muscle mass than the upper limbs, which makes them weigh more. The weight of the lower limbs being greater than the upper limbs may affect reliability in the sense that handling and testing is more difficult.[59, 69]

Table II also presents evidence of larger reliability coefficients when using intraclass correlations than for Cohen's kappas, both with upper and lower extremities and for inter- and intra-rater agreement. In general, the inter-rater reliability appears to be slightly better than the intra-rater one.

Our results also provide evidence that for lower extremities, intraclass correlations exhibited moderate to large heterogeneity, unlike Cohen's kappa which did not. Regarding upper extremities, intraclass correlations also exhibited moderate to large heterogeneity, as well as Cohen's kappas obtained from inter-rater agreement.

Several characteristics of the studies were statistically associated to intraclass correlations both for lower and upper extremities: type of design, study focus, and number of raters. Thus, experimental studies and those with an applied focus presented better reliability than observational and psychometric studies. It is likely that experimental and applied studies have a better control of confounding variables to accurately interpret treatment effectiveness.[31] These results must be interpreted very cautiously due to the small number of experimental studies and also because for upper extremities these two moderator variables reached marginal significance only (P=0.060). In addition, our results found that the larger the number of raters, the lower the reliability. An explanation for this result can be

found in the absence of a standardized protocol when the Modified Ashworth Scale is applied, such as test position, number of repetitions, right-left test order in case of bilateral involvement, testing time (morning or afternoon), etc. When a protocol for administering the Modified Ashworth Scale is missing, then poorer reliability will be expected as the number of raters increases.[47, 52]

The standard deviation of age also showed a statistical association with intraclass correlations for lower extremities, therefore the larger the standard deviation, the lower the reliability. This result has no tentative explanation and must be interpreted cautiously due to the small number of studies on which this result is based (six studies only).

Limitations of the study

Our study has several limitations. First, the small number of studies that reported any reliability estimate with the data at hand limited the generalizability of the results. In particular, the small number of reliability coefficients negatively affected the scope of our analyses of moderator variables. A larger number of studies reporting reliability estimates might have enabled us to propose a predictive model, by means of weighted multiple regression, to at least explain a satisfactory proportion of reliability coefficient's variability. Another limitation was the wide range of reliability coefficients reported in the studies (intraclass correlations, weighted and unweighted Cohen's kappas, squared weighted kappas, Pearson correlations, etc.). This circumstance led us to combine different, but similar, types of agreement coefficients in the same meta-analysis. To normalize the distribution of reliability coefficients and stabilize their variances, Fisher's *Z* transformation was applied for all types of reliability coefficients. From an analytic point of view, instead of transforming kappa coefficients to Fisher's *Z*, it might be advisable to integrate them without using any transformation, but with their own values and the corresponding sampling variance. This strategy was not possible to apply as the studies did not report the statistical data needed to calculate the sampling variance of kappa coefficient.[70] Finally, the poor reporting of several characteristics in the studies limited the possibility of their analysis.

Conclusions

In general, inter- and intra-rater reliability of Modified Ashworth Scale scores was moderate to high. Reliability seemed to be better when measuring upper extremities than lower ones. On average, intraclass correlations seemed to be better than Cohen's kappas, both for inter- and intra-rater agreement. Several study characteristics of the studies were statistically associated to intraclass correlations both for lower and upper extremities: the type of study, the study focus, and the number of raters.

References

1. Gregson JMTJ, Leathley MJ, Moore AP, Smith TL, Sharma AK, Watkins CL. Reliability of measurements of muscle tone and muscle power in stroke patients. Age Ageing 2000;29:223-8.

2. Satkunam LE. Rehabilation medicine: 3. Management of adult spasticity. CMAJ 2003;169:1173-9.

3. Brown P. Pathophysiology of spasticity. J Neurol Neurosurg Psychiatry 1994;57:773-7.

4. Dietz V, Sinkjaer T. Spastic movement disorder: impaired reflex function and altered muscle mechanics. Lancet Neurol 2007;6:725-33.

5. Hsieh JTC, Wolfe DL, Miller WC. A Curt and SCIRE Research Time, Spasticity outcome measures in spinal cord injury: psychometric and clinical utility. Spinal Cord 2008;46:86-95.

6. Morris S. Ashworth and Tardieu scales: their clinical relevance for measuring spasticity in adult and paediatric neurological populations. Phys Ther Rev 2002;7:53-62.

7. Elbasiouny SM, Moroz D, Bakr MM, Mushahvar VK. Management of spasticity after spinal cord injury: current techniques and future directions. Neurorehabil Neural Repair 2010;24:23-33.

8. Francis HP, Wade DT, Turner-Stokes L, Kingswell RS, Dott CS, Coxon EA. Does reducing spasticity translate into functional benefit? An exploratory meta-analysis. J Neurol Neurosurg Psychiatry 2004;75:1547-51.

9. Ada L, O´Dwyer N, O´Neill E. Relation between spasticity, weakness and contracture of the elbow flexors and upper limb activity after stroke: An observational study. Disabil Rehabil 2006;28:891-7.

10. Farmer SE, James M. Contractures in orthopaedic and neurological conditions: A review of causes and treatment. Disabil Rehabil 2001;23:549-58.

11. Barnes MP. Management of spasticity. Age Ageing 1998;27:239-45.

12. Engsberg JR, Ross SA, Collins DR, Park TS. Predicting functional change from preintervention measures in selective dorsal rhizotomy. J Neurosurg 2007;106:282-7.

13. Pandyan AD, Johnson GR, Price CI, Curless RH, Barnes MP, Rodgers H. A review of the properties and limitations of the Ashworth and modified Ashworth Scales as measures of spasticity. Clin Rehabil 1999;13:373-83.

14. Smith AW, Jamshidi M, Lo SK. Clinical measurement of muscle tone using a velocity-corrected modified Asworth scale. Am J Phys Med Rehabil 2002;81:202-6.

15. Mutlu A, Livanelioglu A, Gunel MK. Reliability of Ashworth and Modified Ashworth scales in children with spastic cerebral palsy. BMC Musculoeskelet Disord 2008;9:44.

16. Sampaio C, Ferreira JJ, Pinto AA, Crespo M, Ferro JM, Castro Caldas A. Botulinum toxin type A for the treatment of arm and hand spasticity in stroke patients. Clin Rehabil 1997;11:3-7.

17. Naghdi S, Ansari NN, Azarnia S, Kazemnejad A. Interrater reliability of the Modified Modified Ashworth scale (MMAS) for patients with wrist flexor muscle spasticity. Physioterapy Theory Pract 2008;24:372-9.

18. Annaswamy T, Mallempati S, Allison SC, Abraham LD. Measurement of plantarflexor spasticity in traumatic brain injury: correlational

study of resistance torque compared with the modified Ashworth scale. Am J Phys Med Rehabil 2007;86:404-11.

19. Bohannon RW, Smith MBO. Inter-rater reliability of a Modified Ashworth Scale of muscle spasticity. Phys Ther 1987;67:206-7.

20. Harris JE, Eng JJ. Individuals with the dominant hand affected following stroke demonstrate less impairment than those with the non-dominant hand affected. Neurorehabil Neural Repair 2006;20:380-9.

21. Sloan RL, Sinclair E, Thompson J, Taylor S, Pentland B. Inter-rater reliability of the modified Ashworth Scale for spasticity in hemiplegic patients. Int J Rehabil Res 1992;15:158-61.

22. Klingels K, de Cock P, Molenaers G, Desloovere K, Huenaerts C, Jaspers E, *et al.* Upper limb motor and sensory impairments in children with hemiplegic cerebral palsy. Can they be measured reliably?. Disabil Rehabil 2010;32:409-16.

23. Waninge A, Rook RA, Dijkhuizen A, Gielen E, van der Schans CP. Feasility, test-retest reliability, and interrater reliability of the Modified Ashworth Scale and Modified Tardieu Scale in persons with profound intellectual and multiple disabilities. Res Dev Disabil 2011;32:613-20.

24. Hesse S, Schulte-Tiggees G, Konrad M, Bardeleben A, Wernerr C. Robot-assisted arm trainer for the passive and active practice of bilateral forearm and wrist movements in hemiparetic subjects. Arch Phys Med Rehabil 2003;84:915-20.

25. Hagenbach U, Luz S, Ghafoor N, Berger JM, Grotenhermen F, Brenneisen R, *et al.* The treatment of spasticity with Δ^9 –tetrahydrocannabinol in persons with spinal cord injury. Spinal Cord 2007;45:551-62.

26. Tederko P, Krasuski M, Czech J, Dargiel A, Garwacka-Jodzis I, Wojciechowska A. Reliability of clinical spasticity measurements in patients with cervical spinal cord injury. Ortop Traumat Rehabil 2007;5:467-83.

27. Mutlu A, Livanelioglu A, Gunel Mk. Reliability of Ashworth and Modified Ashworth Scales in Children with Spastic Cerebral Palsy. BMC Musculoskelet Disord 2008;9:14.

28. Paltamaa J, West H, Sarasoja T, Wikström J, Mälkiä E. Reliability of physical functioning measures in ambulatory subjects with MS. Physiother Res Int 2005;10:93-109.

29. Bar-Haim S, Harries N, Copeliovitch L, Ager G, Dobrov I, Kaplanski J. Method of analyzing the performance of self-paced and engine induced cycling in children with cerebral palsy. Disabil Rehabil 2007;29:1261-9.

30. Ansari NN, Naghdi S, Khosravian T, Jalaie S. The interrater and intrarrater reliability of the Modified Ashworth Scale in the assessment of muscle spasticity: Limb and muscle group effect. NeuroRehabilitation 2008;23:231-7.

31. Fosang AL, Galea MP, McCoy AT, Reddihough DS, Story I. Measures of muscle and joint performance in the lower limb of children with cerebral palsy. Dev Med Child Neurol 2003;45:664-70.

32. Yam WKL, Leung MSM. Interrater reliability of modified Ashworth scale and modified Tardieu scale in children with spastic cerebral palsy. J Child Neurol 2006;21:1031-5.

33. Allison SC, Abraman LD. Correlation of quantitative measures with the modified Ashworth scale in the assessment of plantar flexor spasticity in patients with traumatic brain injury. J Neurol 1995;242:699-706.

34. Allison SC, Abraham LD, Petersen CL. Reliability of the Modified Ashworth Scale in the assessment of plantar flexor muscle spasticity in patients with traumatic brain injury. Int J Rehabil Res 1996;19:67-78.

35. Cheng HY, Ju YY, Chen CL, Chuang LL, Cheng CH. Effects of whole body vibration on spasticity and lower extremity function in children with cerebral pasly. Hum Mov Sci 2015;39:65-72.

36. Craven BC, Morris AR. Modified Ashworth scale reliability for measurement of lower extremity spasticity among patients with SCI. Spinal Cord 2010;48:207-13.

37. Streiner DL, Norman GR. Health Measurement scales. A practical guide to their development and use. 4st ed. Oxford University Press: Oxford; 2008.

38. Vacha-Haase T. Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educ Psychol Meas 1998;58:6-20.

39. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Statistics in Medicine 1994;13:2465-76.

40. Sánchez-Meca J, López-López JA, López-Pina JA. Some recommended statistical analytic practices when reliability generalization studies are conducted. Br J Math Stat Psycholm 2013;66:402-25.

41. Duval SJ, Tweedie RL. A non-parametric "trim and fill" method of accounting for publication bias in meta-analysis. JASA 2000;95:89-98.

42. Rothstein HR, Sutton AJ, Borenstein M, editors. Publication bias in meta-analysis: Prevention, assessment and adjustments. Chichester, UK: Wiley; 2005.

43. Borenstein M, Hedges LV, Higgins JPT, Rothstein H. Introduction to meta-analysis. Chichester, UK: Wiley; 2009.

44. López-López JA, Botella J, Sánchez-Meca J, Marín-Martínez F. Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. J Educ Behav Stat 2013;38:443-69.

45. López-López JA, Marín-Martínez F, Sánchez-Meca J, Van den Noortgate W, Viechtbauer W. Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. Br J Math Stat Psychol 2014;67:30-48.

46. Borenstein M, Hedges LV, Higgins JPT, Rothstein H. Comprehensive Meta-analysis 3.3. Englewood Cliffs, NJ: Biostat, Inc; 2014.

47. Kaya T, Karatepe AG, Gunaydin R, Koc A, Altundal-Ercan U. Interrater reliability of the Modified Ashworth Scale and modified Modified Ashworth Scale in assessing poststroke elbow flexor spasticity. Int J Rehabil Res 2011;34:59-64.

48. Numanoglu A, Günel MK. Intraobserver reliability of modified Ashworth scale and modified Tardieu scale in the assessment of spasticity in children with cerebral palsy. Acta Orthop Traumatol Turc 2012;46:196-200.

49. Fasoli SE, Fragala-Pinkham M, Hughes R, Krebs HI, Hogan N, Stein J. Robotic therapy and botulinum toxin type A: a novel intervention approach for cerebral palsy. Am J Phys Med Rehabil 2008;87:1022-5.

50. Kim J, Park HS, Damiano DL. Accuracy and reliability of haptic spasticity assessment using HESS (Haptic Elbow Spasticity Simulator). Conf Proc IEEE Eng Med Biol Soc 2011;2011:8527-30.

51. Salem Y, Lovelace-Chandler V, Zabel RJ, McMillan AG. Effects of prolonged standing on gait in children with spastic cerebral palsy. Phys Occup Ther Pediatr 2010;30:54-65.

52. Gregson JM, Leathley M, Moore AP, Sharma AK, Smith TL, Watkins CL. Reliability of the Tone Assessment Scale and the modified Ashworth scale as clinical tools for assessing poststroke spasticity. Arch Phys Med Rehabil 1999;80:1013-6.

53. Haas BM, Bergström E, Jamous A, Bennie A. The inter rater reliability of the original and of the modified Ashworth scale for the assessment of spasticity in patients with spinal cord injury. Spinal Cord 1996;34:560-4.

54. Mehrholz J, Wagner K, Meissner D, Grundmann K, Zange C, Koch R, *et al.* Relilability of the Modified Tardieu Scale and the Modified Ashworth Scale in adult patients with severe brain injury: a comparison study. Clin Rehabil 2005;19:751-9.

55. Mehrholz J, Major Y, Meissner D, Sandi-Gahun S, Koch R, Pohl M. The influence of contractures and variation in measurement stretching velocity on the reliability of the Modified Ashworth Scale in patients with severe brain injury. Clin Rehabil 2005;19:63-72.

56. Motl RW, Snook EM, Hinkle ML, McAuley E. Effect of acute leg cycling on the soleus H-reflex and modified Ashworth scale scores in individuals with multiple sclerosis. Neurosci Lett 2006;406:289-92.

57. Fasoli SE, Fragala-Pinkham M, Hughes R, Hogan N, Krebs HI, Stein

J. Upper limb robotic therapy for children with hemiplegia. Am J Phys Med Rehabil 2008;87:929-36.

58. Fragala MA, O'Neil ME, Russo KJ, Dumas HM. Impairment, disability, and satisfaction outcomes after lower-extremity botulinum toxin injections for children with cerebral palsy. Pediatr Phys Ther 2002;14:132-44.

59. Clopton N, Dutton J, Featherston T, Grigsby A, Mobley J, Melvin J. Interrater and intrarrater reliability of the Modified Ashworth Scale in children with hyperthonia. Pediatr Phys Ther 2005;17:268-74.

60. Kelly B, Mackay-Lyons MJ, Berryman S, Hyndman J, Wood E. Assessment protocol for serial casting after botulinum toxin a injections to treat equines gait. Pediatr Phys Ther 2008;20:233-41.

61. Li F, Wu Y, Li X. Test-retest reliability and inter-rater reliability of the Modified Tardieu Scale and the Modified Ashworth Scale in hemiplegic patients with stroke. Eur J Phys Rehabil Med 2014;50:9-15.

62. Nunnally JC. Reliability of measurement. In: Mitzel HE, editors. Encyclopedia of educational research. New York: FreePress; 1982. p 1589-601.

63. Campbell SK. On the importance of being earnest about measurement, or, how can be sure that what we know is true? Phys Ther 1987;67:1831-3.

64. Gajdosik RL, Bohannon RW. Clinical measurement of range of motion. Phys Ther 1987;67:1867-72.

65. Vacha-Haase T, Kogan LR, Thompson B. Sample compositions and variabilities in published studies versus those in test manuals. Educ Psychol Meas 2000;60:509-22.

66. Thompson B, ed. Score reliability: Contemporary thinking on reliability issues. Thousand Oaks, CA: Sage; 2003.

67. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304:1491-4.

68. Portney L, Watkins M. Foundations of Clinical Research: Applications to Practice. Second edition. Upper Saddle River, NJ: Prentice Hall Health; 2000.

69. Pandyan AD, Price CIM, Rodgers H, Barnes MP, Johnson GR. Biomechanical examination of a commonly used measure of spasticity. Cin Biomech 2001;16:858-65.

70. Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method 2011;11:145-63.

Conflicts of interest.—The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

Article first published online: September 13, 2017. - Manuscript accepted: September 12, 2017. - Manuscript revised: August 28, 2017. - Manuscript received: April 28, 2017.