SUPPLEMENTARY MATERIAL

- **Appendix 1. Kappa coefficient & Inter-observer agreement in Tinel sign .**

Surgeon_A	Surgeon_B Negative	Surgeon_B Positive	Total
Negative	82	10	92
Positive	1	7	8
Total	83	17	100

A very simple measure of the agreement will be the observer agreement (po = n agreements / total n of observations). The tables in this appendix demonstrate an analysis of a hypothetic data set of the agreement in the assessment of Tinel sign (present vs absent), showing a global agreement of p0 = (82+7)/100 = 89%.

	Estimate	95% Confidence Interval
Specific Agreement (+)	56.0%	37.1% to 73.3% (Wilson)
Specific Agreement (-)	93.7%	89.1% to 96.5% (Wilson)
Global Agreement (+)	89.0%	81.4% to 93.7% (Wilson)
kappa minimum	-0.0582	
maximum	0.7826	

But, is a global agreement of 89% interesting when we do not have agreement in the positive assessment (present Tinel sign)?, Cichetti and Feinstein (1990) recommended the assessment of positive agreement and negative agreement. Positive agreement is the conditional probability of a positive assessment by observer A and observer B (specific positive agreement of p0+= 2*7/(100+7-82)= 56% means that the probability that observer A classified the Tinel sign as positive when observer B classified it as positive too) (see Tables above)). Negative agreement is the conditional probability of a negative assessment by both observers (specific negative agreement of p0- = 2*82/(100+82-7) = 93.7%). Another limitation of the global agreement (p0) is that it does not take into account the agreement that can obtained only by chance, called expected agreement (pe). The maximum true agreement or potential agreement will be equal to 1-pe.

kappa	Standard Error SE0	Standard Error SE1	p value	95% Asymptotic Confidence Interval
0.5063	0.0915	0.1246	0.0000	0.2622 to 0.7504

The Kappa coefficient is one of the most used indexes of agreement which solved that problem because it excluded the expected agreements obtained by chance (Fleiss et al., 2003). It is calculated by dividing the true agreement by the potential agreement (Kappa = p0 - pe / 1 - pe).

Observed Disagreement X-Y+	Observed Disagreement X+Y-	Bias Index	Bias adjusted kappa BAK
1.00%	10.0%	-9.00%	0.4971
Observed Agreement (+)	**Observed Agreement (-)**	**Prevalence Index**	**Prev. & Bias adjusted kappa PABAK**
12.5%	87.5%	-75.0%	0.7800

Other important aspect in the interpretation of the Kappa value is the effect of bias and prevalence. Bias in agreement analysis is the tendency to misclassify variables in a consistent way that tips the scale in one particular direction. In table 1, we can observe how hand surgeon A classified Tinel sign as positive in 10% of the individuals when hand surgeon B classified it as negative. In addition, hand surgeon B classified Tinel sign as positive in 1% of the individuals while hand surgeon A classified it as negative. There was an asymmetric observed disagreement pair. Byrt et al. (1993) defined the Bias Index (BI) as the differences of the proportions of positive assessments between observer A and observer B. The bias-adjusted Kappa (BAK) is the kappa coefficient adjusted for the bias, which is obtained by replacing the disagreement pair by its mean. Prevalence Index is defined as the difference between the proportion of positive and negative assessments by observer A and observer B. The prevalence-adjusted bias-adjusted Kappa (PABAK) is the Kappa coefficient adjusted for bias and prevalence .

- **Appendix 2. Agreement for a categorical variable.**

When the categorical criterion presents more than 2 categories (c>2) with an order, the agreement should be analysed by the weighted kappa coefficient . However, if the categorical response variable is nominal the results of agreement should be based on the unweighted kappa coefficient with a coefficient for each category compared with the rest . An important, and unsolved, problem is that related to the use of the weighted kappa and the number of categories in the response variable.

The following table shows the test - re test reliability analysis in the assessment of Absent Palmaris Longus (PL) in general population. The response variable (Absence of Palmaris Longs) presented 3 categories: No absent PL, Unilateral absent PL, Bilateral absent PL. The results of agreement analysis will be different if the researcher considers the categorical variable nominal or ordinal.

Observat~1	Observation2 Bilateral	Unilateral	No absent	Total
Bilateral	5	0	1	6
Unilateral	0	4	0	4
No absent	1	3	10	14
Total	6	7	11	24

The same observer assessed the palmaris longus (PL) as present vs absent on two occasions, 1 month apart, in a sample of 24 individuals.

A. Nominal categorical response variable (without order)

Weight	Observed Agreement	Kappa	Standard Error SE0	SE1	p value	95% Asymptotic Conf. Interval
UnWeighted	79.2%	0.6648	0.1430	0.1322	0.0000	0.4058 to 0.9238
Categories:						
Bilateral-R	91.7%	0.7778	0.2041	0.1494	0.0001	0.4849 to 1.0707*
Unilateral-R	87.5%	0.6538	0.1915	0.1754	0.0006	0.3101 to 0.9976
No absent-R	79.2%	0.5890	0.1978	0.1586	0.0029	0.2781 to 0.9000

(A) If the response variable is considered as nominal (bilateral absent, unilateral absent or present in both hands) the analysis is done with the unweighted kappa, comparing each class of the response variable with the rest.

B. Ordered categorical response variable

Weight	Observed Agreement	Kappa	Standard Error SE0	SE1	p value	95% Asymptotic Conf. Interval
UnWeighted	79.2%	0.6648	0.1430	0.1322	0.0000	0.4058 to 0.9238
Lineal	85.4%	0.6693	0.1662	0.1424	0.0001	0.3902 to 0.9484
Quadratic	88.5%	0.6733	0.2017	0.1635	0.0008	0.3528 to 0.9937

(B) If the response variable is considered as categorical, in this case with a metric 0,1,2 (0=Bilateral absent, 1=Unilateral absent, 2= Present in both hands) the appropriate agreement coefficient is the weighted kappa.

- **Appendix 3. Intraclass Correlation Coefficient**.

The base of formulation of the ICC was introduced by Fisher (1921) from a special definition of the Pearson correlation coefficient in data with equal mean and variance (Domenech 2017). Actually, the ICC that we use for analysing agreement is based on the analysis of the variance (ANOVA) with repeated measures.

When we face “n” subjects and “K” observers the ICC is based on a two-way ANOVA with two models; random-effect and mixed-effect models. If we have assumed that the n subjects and the K observers came from a random sample of the population, we should use a random-effect model of ICC. Otherwise, if we have considered that the k observers constitute a total representation of the population, we should use a mixed-effect model.

The interpretation of the results differs depending on the model. The ICC obtained by a random-effect model can be generalized to the total population of observers, while the results coming from a mixed-effect model cannot be generalized and it is expected that they will change with a different group of observers. A different situation happens when the assessment has been done by one observer or it is impossible to distinguish the observers, in that case we have to use the “ICC one factor” based on one-way ANOVA, which always presents a random-effect model because the n individuals is a random sample of the population. In summary, for intra-observer reliability (n subjects and one observer) the ICC agreement analysis should be a “one-way, random-effect” model and for inter-observer reliability (n subjects and K observers) could be a “two-way, random-effect” model or “two-way, mixed-effect” model. (Vaz et al. 2013).

One other important aspect of the ICC is that any model (one-way random-effect, two-way random-effect, and two-way mixed-effect) could have two different systems of agreement assessment; consistency and absolute agreement. The following charts help to understand the close concepts of Pearson correlation, ICC consistency agreement, and ICC absolute agreement. Observe the changes in ICCC, ICCA and r related to three different bias situations.. In A the assessments by X and Y are similar and the ICCC, ICCA and Pearson correlation (r) coefficients are similar too. In B it is shown a bias in which one a constant disagreement of -4 between observer X and observer Y. Observer Y assessed, in a systematic way, minus 4 points compared with observer X. The r and ICCC did not detect that bias, only the ICCA showed a lower agreement coefficient of 0.814. In C, it is shown a proportional disagreement in which observer Y assessed in systematic way dividing by 3 the observations done by observer X. In that proportional bias, the r coefficient was unable to detect it, while both the ICCC and ICCA show lower coefficients (0.6 and 0.268 respectively). Chart D shows a constant and proportional disagreement which was detected only by the ICCA (0.147). A constant and proportional disagreement is detected only by ICCA. (based on Domenech 2017). Consequently, for test retest reliability, the recommendation is to use the ICC2,1 that defines a two-way random-effect model with absolute agreement (Vaz et al. 2013; Rosales et al., 2017).

A

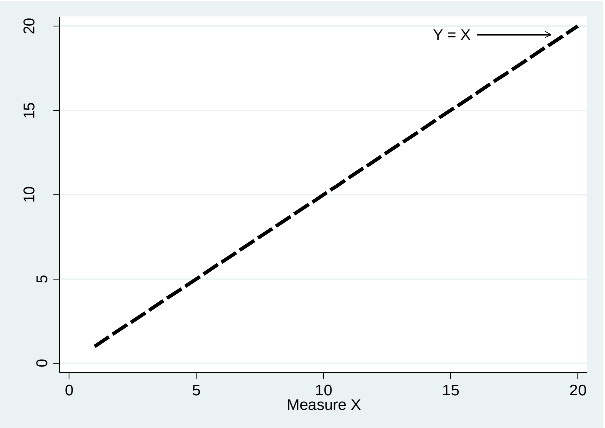
Total Agreement (Y=X)

Consistency: $ICC_C = 1$

Absolute: $ICC_A = 1$

Pearson correlation $r = 1$

Id	X	Y
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
10	10	10



B

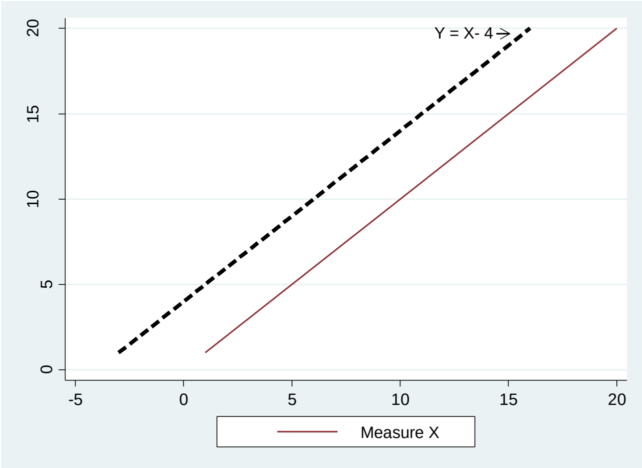
Constant Disagreement (Y=X-4)

Consistency: $ICC_C = 1$

Absolute: $ICC_A = 0.814$

Pearson correlation $r = 1$

Id	X	Y
1	1	-3
2	2	-2
3	3	-1
4	4	0
5	5	1
6	6	2
7	7	3
8	8	4
9	9	5
10	10	6



C

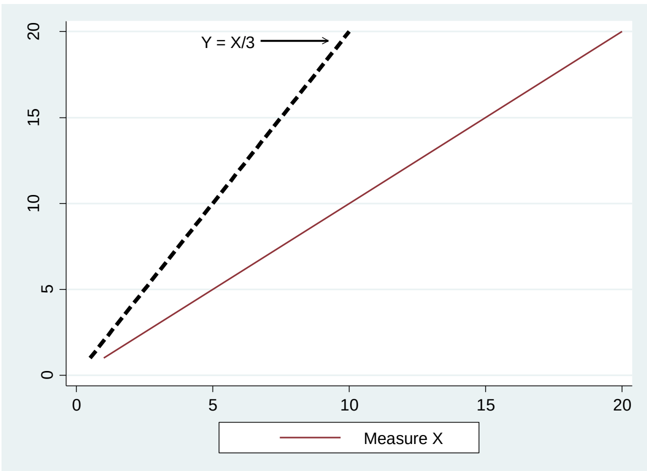
Proportional Disagreement (Y=X/3)

Consistency: $ICC_C = 0.600$

Absolute: $ICC_A = 0.268$

Pearson correlation $r = 1$

Id	X	Y
1	1	.33
2	2	.67
3	3	1
4	4	1.3
5	5	1.7
6	6	2
7	7	2.3
8	8	2.7
9	9	3
10	10	3.3



D

Proportional & Constant

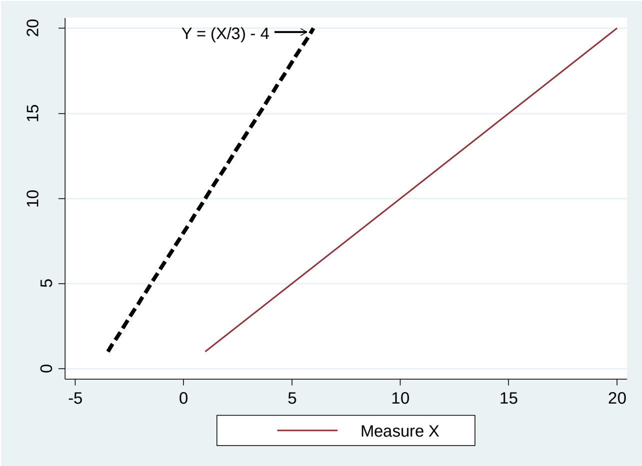
Disagreement (Y=(X/3)- 4))

Consistency: $ICC_C = 0.600$

Absolute: $ICC_A = 0.147$

Pearson correlation $r = 1$

Id	X	Y
1	1	-3.7
2	2	-3.3
3	3	-3
4	4	-2.7
5	5	-2.3
6	6	-2
7	7	-1.7
8	8	-1.3
9	9	-1
10	10	-.67



- **Appendix 4. Bland-Altman Limits of Agreement (LoA) for studying agreement between two methods of assessment of a quantitative criterion.**

The results of comparing two methods (X and Y) of measuring the preoperative pulp-to-pulp pinch strength (Kg) in patients with thumb carpometacarpal (CMC) osteoarthritis.

The Bland-Altman limits of agreement (LoA) calculate the difference between both measures for each subject (di = Yi – Xi) and it is faced with the mean (mi = (Xi + Yi)/2)) of both measures for each individual. If we assume the normality distribution of the differences, it is expected that 95% of the differences should be between the limits of the interval. In fig. 1 we observe how most of the differences are localized between the LoA.

The following tables present the descriptive statistics of Bland-Altman analysis with a listwise strategy which means that the missing values are not included in the analysis.

Bland-Altman: Descriptive Statistics (listwise)

Variable	Valid	Miss	Obs	Mean	Std. Dev.	[95% Conf.	Interval]
Y: P_Pnew	93	0	93	3.030466	.7985896	2.865998	3.194934
X: P_Pstand	93	0	93	3.065689	.7604395	2.909078	3.2223

Valid number of cases (listwise): 93

With a sample of n= 93 subjects and LoA with a 95% CI, it is expected that no more than 4.65 observations (0.05 *93=4.65 or 2.33 over the limit and 2.33 under the limit) will be outside of the LoA. The following table shows that 3 cases were over and 3 cases were under the LoA. Observe that there was no bias (do=- 0.035) with a 95% CI (- 0.1086 to 0.0381637) which was not statistically significant because the 95% CI includes the null hypothesis (H0; do = 0).

Bland-Altman: Absolute values of Bias & Limits of Agreement (LoA)

Parameter		Estimate	Std. Dev.	Std. Err.	[95% Conf.	Interval]
Diff. (Y-X):	Bias	-.0352232	.3563379	.0369505	-.1086101	.0381637
	Lower LoA	-.7336326		.0640002	-.8607424	-.6065227
	Upper LoA	.6631862		.0640002	.5360763	.790296

Cases over limit = 3 (3.23%)
Cases under limit = 3 (3.23%)
Spearman correlation between (Y-X) and (X+Y)/2: r= 0.0964 (p= 0.3579)
Lin's Concordance Correlation coeff. of Absolute Agreement = 0.8947

Besides, some statistics software as in the following table gives the Lin concordance absolute agreement (Lin 1989;1992) which was 0.894 (good to excellent agreement) and the Shapiro-Wilk test for normality (W=0.981, p=0.195) indicating that the differences (d0) followed a normal distribution with information about skewness and kurtosis of the distribution of the differences. The Bland-Altman LoA can be applied only if the differences followed a normal distribution.

Tests of Normality (Y-X)	Statistic	p-value
Shapiro-Wilk	W = 0.9810	0.1952
Skewness	Sk = -0.0437	0.8542
Kurtosis-3	Ku = 1.2039	0.0364
Skewness & Kurtosis	Chi2 = 4.4945	0.1057

Bland-Altman limits of agreement (LoA) method calculates the difference between both measures for each subject (di = Yi – Xi) and it is faced with the mean (mi = (Xi + Yi)/2)) of both measures for each individual. If we assume the normality distribution of the differences, it is expected that the 95% of the differences should be between the limits of the interval. At the end, Stata supplies the normality analysis with the Shapiro-Wilk test (Null hypohesis H0 = There is no difference between the sample and a normal distributed population).

- **Appendix 5**. **Comparison of two methods for measuring pulp-to-pulp pinch strength in patients with thumb CMC osteoartritis by Passing Bablok regression line of agreement**

The Passing and Bablok analysis (Passing and Bablok 1984; 1985; 1988), is a non-parametric estimation of the orthogonal regression line between the two methods. The lineal equation will be Y = A + BX + ε, in which A is the constant difference between the two methods, B is the proportional difference, and ε is a random variable with a mean equal to zero which represents the random non-systematic error between the methods. When A = 0 and B = 1 means that both methods presented the same error and they are comparable. The following table shows an A value -0.1428571 (95% CI: -0.5151515 to 0.0958333)), not significant because the 95% CI includes the H0 = 0. The B coefficient was : 1.047619 (95% CI: 0.9625 to 1.166667), not significant because it included the H0 = 1 (in a rate or proportion H0 = 1 when the numerator is similar to denominator). Consequently, based on the Passing and Bablok analysis we can conclude that the two methods presented no constant and proportional differences and they are comparable. Observe the same Lin´s concordance coefficient of agreement (0.8947) than that one obtained in the Bland-Altman analysis.

```
. agree P_Pnew P_Pstand  , pb

AGREEMENT: PASSING-BABLOK METHOD

Passing-Bablok: Descriptive Statistics (listwise)
----------------------------------------------------------------------------------
Variable     Valid  Miss   Obs    Median       Mean    Minimum    Maximum  Std. Dev.
Y: P_Pnew       93     0    93         3   3.030466   1.333333          5   .7985896
X: P_Pstand     93     0    93  3.090909   3.065689   1.272727   4.818182   .7604395
Y-X             93     0    93  -.030303 -.0352232 -1.075758   1.151515   .3563379
100*(Y-X)/X     93     0    93   -.9009%    -.8411%     -31.9%      36.2%      11.8%
----------------------------------------------------------------------------------
Valid number of cases (listwise): 93

Passing-Bablok: Regression Line (Y = A + B*X)
------------------------------------------------
 A = -.1428571  (95% CI: -.5151515 to  .0958333)
 B =  1.047619  (95% CI:     .9625 to  1.166667)
------------------------------------------------
Linearity Test (CUSUM Test for deviation from linearity):   p > 0.20
Lin's Concordance Correlation coeff. of Absolute Agreement = 0.8947
```

- **Appendix 6. Internal consistency &Cronbach alpha.**

The Cronbach alpha can be derived by using split-half reliability coefficients, calculated by randomly dividing items into two subscales and then correlating the item responses of each subscale with each other. The alpha coefficient represents the average of all possible split-half correlations and typically varies between 0 and 1.

- **Appendix 7. Structural Validity**

The structural validity analysis under CTT is done by Factor Analysis (FA), which is a collection of statistical techniques that try to identify if there are clusters of items that go together. Ideally, a scale measures just one concept, but some concepts may be multidimensional (e.g., PRO instruments that measure health may have one dimension for physical health and a second for mental health). We will expect that those dimensions should be correlated but that correlation should be small enough that we can identify them as two different dimensions of health. Analysing the scale with FA permits us to recognise the set of items that represent each dimension. We may find a first group of items that are strongly correlated and with a pattern of correlation that is consistent, and a second different set of items with a high correlation among them representing the other health dimension, and both sets of items should have a small or moderate correlation with each other.

An alternative method for analysing structural validity is IRT. In CTT, scoring of the measures is usually done by summing or averaging item scores on a set of items, each of which has equal weight. In that way the PRO instrument yields the scores with an error measurement. IRT was developed to overcome the problems of CTT, namely that the item statistics are very dependent on the sample of respondents, the interpretation of respondent characteristics depends on the questionnaire used, and the assumption that errors of measurement are equal for all persons. Finally, CTT cannot make prediction about results for a respondent or a sample, on an item using only their ability scores (Embreston 1966; Embreston & Reise., 2000). ITR enables to predict a person's scores based on his/her abilities or latent traits and to establish a relationship between a person's item performance and the set of traits underling item performance. Consequently, once the data fit an IRT model, the same items may be used in different samples and they will keep their statistical properties (for instance, difficulty and discrimination) independently of the sample's ability to respond to the items (Hamblenton 1996). The use of ITR is not common in PRO assessment in hand surgery but has been used to assess the psychometric properties of the DASH in Dupuytren disease (Forget et al. , 2014) and the CTS-6 in CTS (Lyren et al 2012).

- **Appendix 8. References (complete list)**

    - Alonso J, Prieto L, Antó JM. [The Spanish version of the SF-36 Health Survey (the SF-36 health questionnaire): an instrument for measuring clinical results]. Med Clin (Barc). 1995, 27;104:771-6
    - Amadio P.C. Outcome assessment in hand surgery. Clin Plast Surg, 1997; 24: 191-4.
    - An TW, Evanoff BA, Boyer MI, Osei DA. The Prevalence of Cubital Tunnel Syndrome: A Cross-Sectional Study in a U.S. Metropolitan Cohort. J Bone Joint Surg Am 2017;99:408–416ç
    - Arias-de la Torre J, Puigdomenech E, Valderas JM, Evans JP, Martín V, Molina AJ, Rodríguez N, Espallargues M. Availability of specific tools to assess patient reported outcomes in hip arthroplasty in Spain. Identifying the best candidates to incorporate in an arthroplasty register. A systematic review and standardized assessment. PLoS One. 2019 Apr 1;14(4):e0214746
    - Atroshi I, Johnsson R, Sprinchorn A. Self-administered outcome instrument in carpal tunnel syndrome: Reliability, validity and responsiveness evaluated in 102 patients. Acta Orthop Scand 1998; 69 : 82-88.
    - Atroshi I, Gummesson C, Johnsson R, Ornstein E, Ranstam J, Rosén I. Prevalence of carpal tunnel syndrome in a general population. JAMA 1999;282:153–158
    - Atroshi I, Gummesson C, Johnsson R, McCabe SJ, Ornstein E. Severe carpal tunnel syndrome potentially needing surgical treatment in a general population. J Hand Surg Am. 2003; 28:639-44.
    - Atroshi I, Lyrén PE, Gummesson C. The 6-item CTS symptoms scale: a brief outcomes measure for carpal tunnel syndrome. Qual Life Res. 2009 ;18:347-58
    - Bablok W, Passing H. Application of statistical procedures in analytical instrument testing. J Automat Chem. 1985;7:74-9
    - Bablok W, Passing H, Bender R, Schneider B. A general regression procedure for method transformation. Application of linear regression procedures for method

comparison studies in clinical chemistry, Part III. J Clin Chem Clin Biochem. 1988; 26:783-90

- Bayes T. An essay towards solving a problem in the doctrine of chances. 1763. MDComput 1991; 08:157–171
- Beaton DE, Katz JN, Fossel AH, et al. Validity, reliability and responsiveness of the DASH outcome measure in different regions of the upper extremity. Journal of Hand Therapy, 2001; 14: 1204-17.
- Beaton DE, Wright JG, Katz JN; Upper Extremity Collaborative Group. Development of the QuickDASH: comparison of three item-reduction approaches. J Bone Joint Surg Am. 2005 ;87:1038-46
- Bravo G, Potwin L. Estimating the reliability of continuos measures with Cronbach´s alpha or the intraclass correlation coefficient: toward the integration of two traditions. J Clin Epidemiol 1991; 44: 381-90.
- Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1981; 1: 307-310.
- Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol, 1993; 46: 423-9
- Cicchetti DV, Feinstein AR. High agreement but low Kappa: II. Resolving the paradoxes. J Clin Epidemiol, 1990; 43:551-8.
- Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968;70:213-20.
- Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297-334.
- Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health related quality of life. J Clin Epidemiol 2003; 56: 395–407.
- de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006 Aug 22;4:54
- de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, Boers M, Bouter LM. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res. 2007 ;16:131-42.
- Domenech JM. Fundamentos de Diseño y Estadistica. 18th ed. Barcelona, Signo, 2017
- Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16 Suppl 1:5-18.
- Embretson SE. The new rules of measurement. Psychol Assess. 1996;8(4):341–9.
- Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.; 2000.
- Evans JP, Porter I, Gangannagaripalli JB, Bramwell C, Davey A, Smith CD, Fine N, Goodwin VA, Valderas JM. Assessing Patient-Centred Outcomes in Lateral Elbow Tendinopathy: A Systematic Review and Standardised Comparison of English Language Clinical Rating Systems. Sports Med Open. 2019 Mar 20;5(1):10

- Forget NJ, Jerosch-Herold C, Shepstone L, Higgins J. Psychometric evaluation of the Disabilities of the Arm, Shoulder and Hand (DASH) with Dupuytren's contracture: validity evidence using Rasch modeling. BMC Musculoskelet Disord. 2014 Oct 30;15:361.
- Fisher RA. On the “probable error” of a coefficient of correlation deduced from a small simple. Metrom, 1921; 1: 1-32.
- Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Edu Psychol Meas 1973; 33: 613-9.
- Fleiss JL, Levin B, Paink MC. Statistical Methods for Rates and Proportions. 3rd ed. Hoboken NJ. John Wiley & Sons, 2003: 598:608.
- Gummenson C, Atroshi I, Ekdahl C. The quality of reporting and outcome measures in randomized clinical trials related to upper-extremity disorders. J Hand Surg Am, 2004; 29: 727-34.
- Guyatt GH. A taxonomy of health status instruments. J Rheumatol 1995;22:1188–1190.
- Hambleton RK. Item response theory: A broad psychometric framework for measurement advances. Psicothema. 1994;6:535–56.
- Harden RN, Bruehl S, Perez RS, et al. Validation of proposed diagnostic criteria (the "Budapest Criteria") for Complex Regional Pain Syndrome. Pain. 2010 ;150:268-74.
- Hoekstra CJ, Deppeler DA, Rutt RA. Criterion validity, reliability and clinical responsiveness of the CareConnections Functional Index. Physiother Theory Pract. 2014 ;30: 429-37.
- Jaeschke R, Singer J, Guyatt GH. Ascertaining the minimal clinically important difference. Cont Clin Trials 1989;10:407-415.
- Kennedy CA, Beaton DE, Solway S, McConnell S, Bombardier C. The DASH and Quick DASH outcome measure user´s manual. Third Edition. Toronto, Ontario: Institute for Work & Health, 2011.
- Landis JR, Koch GG. The measurements of observer agreement for categorical data. Biometrics 1977, 33: 159-74.
- Lawshe CH. A quantitative approach to content validity. Personnel Psychol, 1975; 28: 563-575.
- Lyrén PE, Atroshi I. Using item response theory improved responsiveness of patient-reported outcomes measures in carpal tunnel syndrome. J Clin Epidemiol. 2012; 65:325-34.
- Liang MH. Evaluating Measurement Responsiveness. J Rheumtol, 1995; 22 : 1191-1192.
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996; 1: 30-46.
- Mohan A, Vadher J, Ismail H, Warwick D. The Southampton Dupuytren's Scoring Scheme. J Plast Surg Hand Surg. 2014 ;48:28-33.
- Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010 , 18;10:22.

- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patientreport outcomes. Journal of Clinical Epidemiology. 2010;63:737-45.
- Ozyürekoğlu T, McCabe SJ, Goldsmith LJ, LaJoie AS. The minimal clinically important difference of the Carpal Tunnel Syndrome Symptom Severity Scale. J Hand Surg Am. 2006 ;31:733-8
- Page RM, Cole GE, Timmreck TC. Basic Epidemiological Methods and Biostatistics. A practical guide book. Suddbury MA, Jones & Barret Publishers, 1995: 146-147
- Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part II. J Clin Chem Clin Biochem. 1984 ;22:431-45.
- Petersen MA, Groenvold M, Bjorner JB, Aaronson NK, Conroy T, Cull A, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res 2003;12:373-85.
- Rodrigues JN, Mabvuure NT, Nikkhah D, Shariff Z, Davis TR. Minimal important changes and differences in elective hand surgery. J Hand Surg Eur . 2015;40:900-12.
- Rodrigues J, Zhang W, Scammell B, Russell P, et al.. Validity of the Disabilities of the Arm, Shoulder and Hand patient-reported outcome measure (DASH) and the Quickdash when used in Dupuytren's disease. J Hand Surg Eur Vol. 2016 l;41:589-99.
- Rosales RS, Reboso-Morales L, Martin-Hidalgo Y, Diez de la Lastra-Bosch I. Level of evidence in hand surgery. BMC Res Notes. 2012, 2;5:665.
- Rosales RS. Clinical research in hand surgery. J Hand Surg Eur , 2015 ;40:546-8.
- Rosales RS, Atroshi I. Basics of Statistics for Clinical Research in Hand Surgery. Rev Iberam Cir Mano 2018;46:141–16.
- Rosales RS, Delgado EB, Díez de la Lastra-Bosch I. Evaluation of the Spanish version of the DASH and carpal tunnel syndrome health-related quality-of-life instruments: cross-cultural adaptation process and reliability. J Hand Surg Am. 2002, 27:334-43.
- Rosales RS, Diez de la Lastra I, McCabe S, Ortega Martinez JI, Hidalgo YM. The relative responsiveness and construct validity of the Spanish version of the DASH instrument for outcomes assessment in open carpal tunnel release. J Hand Surg Eur, 2009; 34:72-5
- Rosales RS, Atroshi I. Spanish versions of the 6-item carpal tunnel síndrome symptoms scale (CTS-6) and palmar pain scale. J Hand Surg Eur , 2013 ;38:550-51.
- Rosales RS, Martin-Hidalgo Y, Reboso-Morales L, Atroshi I. Reliability and construct validity of the Spanish version of the 6-item CTS symptoms scale for outcomes assessment in carpal tunnel syndrome. BMC Musculoskelet Disord. 2016 , 3;17:115.
- Rosales RS, García-Gutierrez R, Reboso-Morales L, Atroshi I. The Spanish version of the Patient-Rated Wrist Evaluation outcome measure: cross-cultural adaptation process, reliability, measurement error and construct validity. Health Qual Life Outcomes. 2017, 24;15(1):169.

- Sandroni P, Benrud-Larson LM, McClelland RL, Low PA. Complex regional pain syndrome type I: incidence and prevalence in Olmsted county, a population-based study. Pain. 2003;103:199-207.
- Silman AJ. Epidemiological Studies: a practical guide. New York: Cambridge University Press; 1995.
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420-8 .
- Shultz KS, Whitney DJ, Zickar MJ. Measurement theory in action: Case studies and exercises. 2nd ed. New York, Routledge, 2014.
- Streiner DL, Norman GR. Health measurement scales. A practical guide to theirdevelopment and use. 4 ed. New York: Oxford University Press; 2008.
- Stratford PW, Gill C, Westaway M, Binkley JM. Assessing disability and change on individual patients: a report of a patient-specific measure. Physiother Can 1995;47:258–63.
- Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012, 21: 651-7.
- Valderas JM, Ferrer M, Mendívil J, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health 2008;11:700–8.
- Van Abswoude AAH, Van der Ark LA, Sijtsma K. A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models. Applied Psychology Measurement 2004;28:3-24.
- Vaz S, Falkmer T, Passmore AE, Parsons R, Andreou P. The case for using the repeatability coefficient when calculating test-retest reliability. PLoS One. 2013; 8(9):e73990.
- Vrotsou K, Ávila M, Machón M, Mateo-Abad M, Pardo Y, Garin O, Zaror C, González N, Escobar A, Cuéllar R. Constant-Murley Score: systematic review and standardized evaluation in different shoulder pathologies. Qual Life Res. 2018 ;27:2217-2226
- Ware JE Jr, Gandek BL, Keller SD. The IQOLA Project Group. Evaluating instruments used cross-nationally: methods from the IQOLA project. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed.Philadelphia: Lippincott-Raven, 1996:337–346
- Wilcke MT, Abbaszadegan H, Adolphson P Y . Evaluation of a Swedish version of the patient-rated wrist evaluation outcome questionnaire: good responsiveness, validity, and reliability, in 99 patients recovering from a fracture of the distal radius. Scand.J Plast.Reconstr.Surg.Hand Surg, 2009; 43: 94-101.
- World Health Organization. How to use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for comment. . Geneva: WHO , 2013.