Original Article

Systematic review of the measurement properties of instruments utilized to diagnose Temporomandibular Disorders according to the COSMIN checklist

Marília Barbosa Santos Garcia,[1] Ana Paula Amaral,[1] Cid Andre Fidelis de Paula Gomes,[2] Fabiano Politti,[1] Daniela Aparecida Biasotto-Gonzalez,[1]

[1]Postgraduate Program in Rehabilitation Sciences, Núcleo de Apoio a Pesquisa em Análise do Movimento, Universidade Nove de Julho (UNINOVE), Rua Profa Maria Jose Barone Fernandes, 300, São Paulo, SP 02117-020, Brazil. mariliabsg@gmail.com; ap.fisioterapeuta@gmail.com; politti@uninove.br; dani_atm@uninove.br

[2]Postgraduate Program in Biophotonics Applied to Health Sciences, Universidade Nove de Julho (UNINOVE), Rua Vergueiro, 235, São Paulo, SP, 01504-001, Brazil. cid.andre@gmail.com

Abstract

Introduction: Most of the instruments used in Brazil to diagnose Temporomandibular Disorders (TMD) were developed in another Language. To effectively use instruments that were created in another language, it must be translated into the relevant target language before cross-cultural adaptation. Clinimetric tests should also be performed. Measurement properties consist of quality criteria related to evaluation instruments. These criteria are necessary to determine the quality of the instruments used in Brazil. Objectives: The aim of the present systematic review was to assess the quality of the measurement properties of instruments utilized to diagnose TMD. **Methods:** Systematic searches were performed of the PUBMED, SCIELO, LILACS and SCIENCE DIRECT databases. Studies addressing questionnaires translated and cross-culturally adapted for use in Brazil were retrieved and the quality of the measurement properties was analyzed using the COSMIN checklist. **Results:** In the 11 eligible articles, 10 instruments were identified. The studies were evaluated based on their analysis of structural validity, internal consistency, reproducibility (concordance and reliability), responsiveness, ceiling effect and floor effect. None the assessment tools had all its measurement properties tested. **Conclusion:** The measurement properties of the instruments were not completely tested. Thus, care must be taken when interpreting the scores of these questionnaires.

Keywords: Temporomandibular disorder, Measurement Properties, Questionnaire.

Correspondence to:
Marília Barbosa Santos Garcia
Rua Andrea Del Sarto, 205, Jardim Martineli, Itatiaia
CEP: 27498000, RJ, Brazil
Telephone: +55 24988332338
Email: mariliabsg@gmail.com

Introduction

Temporomandibular disorder (TMD) regards a set of clinical problems involving articular noises, limited range of motion and/or deviations in mandibular function, pre auricular pain, as well as pain in the temporomandibular joint (TMJ) and/or

masticatory muscles[1].

The literature offers different instruments to diagnose TMD, with distinct categorizations: questionnaires, anamnestic indices, clinical indices and diagnostic criteria[2-7]. Most of the instruments used in Brazil to diagnose Temporomandibular Disorders (TMD) were developed in another Language. To effectively use instruments that were created in another language, it must be translated into the relevant target language before cultural adaptation. Clinimetric tests should also be performed to evaluate the measurement properties. This procedure is fundamental due to the different customs, cultures, languages and perceptions of health found in different countries. The culturally adapted questionnaire overcomes linguistic and cultural barriers[8].

That assessment instruments are only useful and capable of providing scientifically robust results when they demonstrate satisfactory measurement properties, that is, that all the measurement properties have been tested with adequate sample and also that they present values statistically indicated by the criterion of quality followed when performing the clinimetric tests[8].

Despite the significant increase in the quantity of assessment scales and/or questionnaires, many of them have not been developed and/or validated appropriately[9].

Studies that evaluate the measurement properties of assessment tools should have a high degree of methodological quality. To evaluate the quality of such studies, criteria are needed to classify the study design and statistical analyses. The Consensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist provides these criteria[10-14]. This list was developed in an international, multidisciplinary study involving the participation of 43 specialists in measurement properties in the field of health[10-14]. According to the COSMIN group, studies assessing measurement properties should exhibit a high methodological quality in order to ensure appropriate conclusions concerning the validity of the instrument[10-14].

The aim of the present systematic review was to employ the COSMIN checklist to analyze the methodological quality of measurement properties of TMD assessment tools for use in Brazil.

Methods

The present study was a systematic review, which followed the recommendations of the PRISMA checklist. It was registered under number 2014 CRD42014014286 in PROSPERO (International prospec tive register of systematic reviews) and can be accessed at http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID = CRD42014014286. The details of the protocol of this systematic review can be accessed using the following link: http://www.crd.york.ac.uk/PROSPEROFILES/14286_PROTOCOL_20140920.pdf.

Study selection

Systematic searches were performed of the PUBMED, SCIELO, LILACS and SCIENCE DIRECT databases. The search terms and operators (AND, OR or NOT) used in the electronic databases were as follows: *"temporomandibular disorder"* or *"temporomandibular dysfunction"* or *"temporomandibular*

pain" and *"questionnaire"* or *"index"* or *"scale"* or *"score"* or *"assessment"* or *"evaluation"* or *"self-report"* or *"inventory"* and *"Brasil"* or *"Brazil"* or *"Portuguese"* or *"Brazilian-Portuguese"* or *"Brazilian".* No restrictions were imposed regarding language or publication date. The last search was performed on Dez 17th, 2015.

Inclusion criteria

The study included instruments used in Brazil to diagnose TMD that had been submitted to a procedure for testing the measurement properties. This study included instruments with variable characteristics of the sample (with TMD, without TMD, or both), adult or pediatric population and different statistics analyses. Only studies that were published in complete texts were included.

Exclusion criteria

Texts that were part of theses and dissertation, congress summaries or books were excluded.

Data extraction and assessment of the methodological quality of the eligible studies

Data referring to measurement properties were extracted from each study and analyzed according to the COSMIN checklist[10-14].

When assessing the quality of an instrument, one can distinguish three domains of quality: reliability; validity and responsiveness. Each domain contains one or more measurement properties. The domain reliability contains three measurement properties: internal consistency; reliability and measurement error. The domain validity also contains three measurement properties: content validity; structural validity and criterion validity. The domain responsiveness contains only one measurement property, also known as responsiveness. A number of measurement properties contain one or more aspects, which were defined separately: validity of the content included face validity; structural validity includes structural validity, tests of hypotheses and cross-cultural validity[10-14].

According to COSMIN, the main measurement properties are the following: internal consistency; validity; reliability; the measurement error and ceiling and floor effects. Internal consistency is a measurement of the homogeneity of an instrument and indicates the degree to which the items of a determined instrument are correlated, thereby measuring the same construct. Validity indicates whether the instrument is assessing the construct it proposes to measure and can be used to measure the criterion validity (in the case of a "gold standard") or the structural validity (when there is no "gold standard" for comparison). Reliability refers to the capacity of a certain test to obtain similar results for stable individuals. The measurement error confirms the errors in patients scores that did not attribute real changes to the construct that was measured. Ceiling and floor effects refer to the number of individuals interviewed that reached the maximum and minimum score possible, respectively[10-14]. The COSMIN checklist has 12 boxes, ten of which can be used to evaluate whether a study meets the requirements of adequate methodological quality. Nine of these boxes have norms for measurement properties: Box A (internal consistency); Box B (reliability); Box C (measurement error); Box D (content validity, including face validity); Box E (structural

validity), Box F (hypotheses testing); Box G (cross-cultural), Box H (criterion validity) and Box I (responsiveness). Box J is used for the interpretability of the results of a given study. Box IRT is used for Item Response Theory and the Generalizability Box is used for the possibility of the generalization of a study regarding one or more measurement properties[13].

Part of the COSMIN group developed an evaluation scale to classify each measurement property as excellent, good, reasonable or weak based on the scores of the items in the corresponding box. Methodological quality using this scale is defined by the worst score of a given box. Thus, a box with some items classified as excellent or good, but one item classified as poor is classified as having poor methodological quality ("worse score counts")[14].

The extraction of the data and evaluations were performed by a single rater and verified by an independent reviewer, who then met to discuss the findings. No divergences of opinion were found between the rater and independent reviewer.

Results

In total, 513 studies were found in the searches, although only 11 were considered eligible for the data analysis (Figure 1).
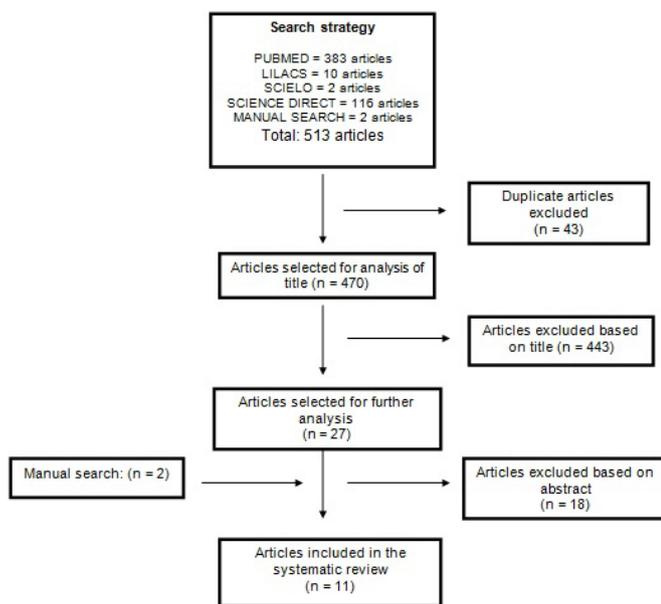

Fig. 1 - Flowchart of systematic review.

In the 11 eligible articles, ten instruments were identified: the Research Diagnostic Criteria for Temporomandibular Disorders(RDC/TMD)[15,16]; the multimedia version of the RDC/TMD[17]; the Mandibular Function Impairment Questionnaire (MFIQ)[18]; the Fonseca anamnestic index (FAI)[19]; the category-ratio scale (CR-10)[20]; the reduced version of the FAI[21]; the self-report for oral conditions[22]; the Brazilian version of the Manchester Orofacial Pain Disability Scale (Brasil-MOPDS)[23]; a screening questionnaire for orofacial pain and temporomandibular disorders, recommended by the American Academy of Orofacial Pain (QAADO)[24]; and the

reduced version of the QAADO[25]. A number of the measurement properties from all of the instruments found in the searches were tested.

The table 1 displays the sample size, sample description and statistical values of each measurement property assessed in the studies included in this systematic review. The table 2 displays the assessments of measurement properties, according to the COSMIN checklist[10-14].

In summary, the measurement properties reliability, internal consistency and content validity were tested for the RDC/TMD[15,16]. Reliability and internal consistency were tested for the multimedia version of the RDC/TMD[17]. Reliability, internal consistency and the content validity were tested for the Mandibular Function Impairment Questionnaire (MFIQ)[18]. Internal consistency and reliability were analyzed for the FAI[19]. Only the content validity was analyzed in the CR-10 for TMD[20]. Internal consistency, reliability and validity were tested for the reduced version of the FAI[21]. Internal consistency, reliability and the criterion validity were analyzed for the self-report of oral conditions[22].Internal consistency, reliability and the structural validity were tested for the Brasil-MOPDS[23]. Only reliability was tested for the QAADO[24], whereas reliability, internal consistency and the content validity were tested for the reduced version of the QAADO[25].

Lucena et al.[15] analyzed the internal consistency (Box A), reliability (Box B) and structural validity (Box E) of the RDC/TMD, for which the respective classifications were good, fair (based on the moderate sample size [n = 45], although other items were classified as good and excellent) and poor. The poor classification for structural validity was due to an error regarding the formulation of hypotheses, which were not described prior to testing the validity. Campos et al.[16] tested the internal consistency (Box A) and reliability (Box B) of the RDC/TMD, for which the respective classifications were good and fair (due to the sample size [n = 36]).

Campos et al.[19] analyzed the internal consistency (Box A) and reliability (Box B) of the Fonseca Index, for which the respective classifications good (due to the failure to perform factor analysis) and fair (due to the moderate sample size [n = 40]). Ferreira-Bacci et al.[20] analyzed the content validity (Box D) of the CR-10 questionnaire used to measure pain associated with TMD, for which the classification was fair. Cavalcanti et al.[17] analyzed the internal consistency (Box A) and reliability (Box B) of the multimedia version of the RDC/TMD, for which both classifications were fair due to the failure to perform factor analysis and the moderate sample size (n = 30).

Campos et al.[18] tested the measurement properties of the MFIQ. Internal consistency (Box A) was classified as good, with some items were classified as excellent, such as the use of factor analysis. Reliability (Box B) was classified as good due to the adequate sample size (n = 62). Content validity (Box D, including face validity) was classified as excellent. This was the only study to employ Cronbach's alpha coefficient in combination with factor analysis, which is an important analysis, as it allows the identification of subscales on a questionnaire and Cronbach's alpha coefficient should be calculated separately for each subscale[26]. The studies that evaluated the RDC/TMD and the multimedia version of the RDC/TMD calculated Cronbach's alpha coefficients, but did not employ factor analysis.

Table 1 - Classification of the measurement properties of the articles included in the review according to the COSMIN checklist.

Study	Sample Size	Sample Description	Reliability	Internal Consistency	Responsiveness	Validity
Lucena et al., 2006 (RDC/TMD)	n=155 n=45 (retest)	Volunteers with TMD	Kappa between 0,73 and 0,91 (p<0,01)	Cronbach alpha value= 0,72	_	Spearman correlation between the OHIP14 and OHIP 0.598 0.349 0.336 0.261
Campos et al., 2007 (RDC/TMD)	n=109 n=36 (retest)	Volunteers with TMD	Intra-examiner: ICC- Questions: 4a = 0, 9869 4b = 0, 9856 7 = 0, 8302 8 = 0, 9869 9 = 0, 9661 11 = 0, 9850 12 = 0, 9785 13 = 0, 9265	Cronbach alpha value= 0, 8479 (intensity of pain and disability) 0, 8971 (limitation of mandibular function) 0, 8673 (non-specific physical symptoms including pain) 0, 8080 (non-specific physical symptoms excluding pain) 0, 9270 (depression)	_	_
Ferreira-Bacci et al., 2009 (CR-10)	n=121	Volunteers with TMD	_	_	_	Pearson correlation between the VAS (p=0,76)
Campos et al., 2009 (FAI)	n=1230 n=40 (retest)	Volunteers without TMD	Kappa Question1=0.787 Question 2=0.725 Question 3=0.771 Question 6=0.805 Question 7=0.838	Kr-20= 0.5594	_	_
Cavalcanti et al., 2010 RDC/TMD multimedia version)	n=30 (15 responded in writing and 15 responded to the multimedia version)	Volunteers with TMD	_	Cronbach alpha value= 0.94	_	Spearman correlation with the RDC/TMD written version 0.670 to 0, 913 p < 0.01
Campos et al., 2012 (MFIQ)	n=62(responded twice – reproducibility) n=219 (validation)	Volunteers with TMD	ICCD1= 0,895, 95% IC D1= 0,832 to 0,935 ICCD2= 0,825, 95% IC D2= 0, 726 to 0, 891	Cronbach alpha valueD1= 0, 874 Cronbach alpha value D2= 0, 918	_	Content validity (face validity) CVR minimum of 0.43
Manfredri et al., 2001 (QAADO)	n=46	Volunteers with TMD	_	k= 0, 454 (muscle disorders) k= 0, 043 (intra-joint disorders	_	_
Kallás et al., 2012 (Brasil-MOPDS)	n=50	Volunteers with orofacial pain and TMD	ICC (inter-examiner)= 0.92 ICC (intra-examiner)= 0.98	Cronbach alpha value= 0.9	_	Structural validity (Pearson) with OHIP14 (r= 0.85) and with VAS (r= 0.75)
Franco-Micheloni et al., 2014 (QAADO reduced version)	n=1307	Volunteers with and without TMD	K=0.840	Kr-20= 0.673	_	Content validity CVR= 0.42
Campos et al., 2014 (FAI – reduced version)	n=700	Volunteers without TMD	_	Cronbach alpha value= 0.745	_	Structural validity AVE = 0.513, CR = 0.878 Concurrent validity with the MIFQ r=0.66, p<0.01
Pinelli and Loffredo, 2007 (self-report of oral conditions)	n=200	Volunteers with orofacial pain	(k = 0.85 for the condition of the TMJ) (k = 0.81 for the periodontal condition)and (k = 0.69 for the dental condition)	_	_	Criterion validity with clinical examinations of the WHO manual

Legend: k =Kappa; ICC = intraclass correlation coefficient; CI = confidence interval; Kr-20 =Kuder-Richardson reliability coefficient; CVR= content validity ratio; D1= functional capacity dimension; D2= nourishment dimension; ES= effect size; AVE= average variance extracted; CR = composite reliability; WHO= world health organization; OHIP14= Oral Health Impact Profile-14; OHIP = Oral Health Impact Profile. VAS = Visual Analogue Scale; MIFQ= Mandibular Function Impairment Questionnaire.

Table 2 - Classification of questionnaires related to TMD according to the COSMIN checklist..

Studies	Measurement property analyzed	Classification according to the COSMIN checklist
Lucena et al., 2006 (RDC/TMD)	Internal consistency	Good
	Reliability	Reasonable
	Structural validity	Weak
Cavalcanti et al., 2010 (RDC/TMD)	Internal consistency	Reasonable
	Reliability	Weak
Campos et al., 2007 (RDC/TMD)	Internal consistency	Good
	Reliability	Reasonable
Ferreira- Bacci et al., 2009 (CR-10)	Structural validity	Weak
Campos et al., 2012 (MIFQ)	Internal consistency	Good
	Reliability	Reasonable
	Content validity	Good
Manfredi et al., 2001 (QAADO)	Reliability	Weak
Campos et al., 2009 (FON-SECA)	Internal consistency	Good
	Reliability	Reasonable
Kallás et al., 2012 (Brazil-MOPDS)	Internal consistency	Good
	Reliability	Reasonable
	Structural validity	Good
Franco-Micheloni et al., 2014 (QAADO reduced version)	Internal consistency	Good
	Reliability	Reasonable
	Content Validity	Good
Campos et al., 2014 (FAI, reduced version)	Internal consistency	Good
	Structural validity	Good
Pinelli & Loffredo, 2007 (self-report of oral conditions)	Internal consistency	Weak
	Reliability	Reasonable
	Criterion validity	Reasonable

Manfredi et al.[24] analyzed the reliability (Box B of the COSMIN checklist) of the QAADO. Although the sample size (n = 46) was classified as fair and other items were classified as good, the study was classified as poor with regard to reliability due to the failure to calculate the intraclass correlation coefficient (ICC), since the "worst score counts" on the COSMIN checklist[14].

Only the studies analyzing the RDC/TMD[15] and MIFQ[18] mentioned the type of Intraclass Correlation Coefficient (ICC) employed to measure reliability. Both studies used 95%. It is extremely important to state the type of ICC employed, as different ICCs can demonstrate completely different results, which can either underestimate or overestimate reliability[27].

The study that analyzed the Brazil-MOPDS[23] questionnaire tested the following measurement properties: internal consistency (box A), reliability (box B) and structural validity (box E). The internal consistency obtained a good classification. The reliability reached a reasonable classification because it did not present the type of ICC used in the analysis. And the structural validity was classified as good.

The study developed by Franco-Micheloni et al.[25] tested the following measurement properties of QAADO instruments reduced version: internal consistency (box A), reliability (box B) and construct validity (box E). The internal consistency has been classified as good due to the factorial analysis. The reliability obtained a reasonable classification because it did not present the type of ICC used in the analysis. Finally, the construct validity presented a good methodological quality.

The self-report of oral conditions had the properties of measure reliability and internal consistency tested through the study of Pinelli and Loffredo[22].The reability was classified as reasonable because it did not use ICC in statistical analysis. The internal consistency obtained a weak degree because Cronbach's alpha was not calculated. The criterion validity (Box E) of the self-report of oral condition was also tested. A reasonable classification was reached because it did not use a gold standard questionnaire in the statistical comparison.

Discussion

The aim of the present study was to assess the quality of the measurement properties of instruments to diagnose temporomandibular disorder that had been translated to Portuguese.

None of the questionnaires completely tested all measurement properties. Reliability was tested in 81.8% of the questionnaires, with 66.7% classified as reasonable. This classification was mainly due to the fact that none of the articles mentioned which type of Intraclass Correlation Coefficient (ICC) was used to measure reliability. It is extremely important to specify which type of ICC was used in different tests, given that different ICCs can lead to completely different results, which would underestimate or overestimate the reliability, depending on the ICC used[27].

Internal consistency was also assessed in 81.8% of the instruments, with 77.8% receiving a classification of good. Questionnaires did not receive a classification of excellent if they did not use the Cronbach alpha test in combination with factorial analysis. This analysis is important since it can identify how many scales are present in a questionnaire. If there is more than one scale, the Cronbach alpha value should be calculated for each sub-scale separately[26]. Only the reduced version of the FAI and the MIFQ used the Cronbach alpha test in combination with factorial analysis.

Measurement errors were not tested in any of the questionnaires. The measurement error confirms errors in the scores of patients that did not attribute real changes in the construct that was measured[14].

The criterion validity was tested in 9.09% of the instruments and was classified as reasonable. The content validity was analyzed in 18.9% of the studies and was classified as good. The structural validity was tested in 46.1% of the instruments. In these studies, there were no classifications of excellent. The studies used Pearson's correlation test to correlate a specific questionnaire with other similar measurements. However, prior to testing the structural validity, it is important to formulate hypotheses that should specify the range and the direction of the expected correlation. Hypotheses were not formulated in any of the studies included in this review. Without specific hypotheses, the risk of bias can be high since it is easier to develop an alternative explanation for low correlations, rather than concluding that the questionnaire may not exhibit high indices for the validity of the construct[26].

Responsiveness was not assessed in any of the studies included in this systematic review. Responsiveness represents the ability of a questionnaire to detect clinical changes over time[14]. In addition, none of the studies tested ceiling/floor effects and consequently, it is not clear if the instruments assessed would fail to detect an improvement or a worsening in certain patients.

Costa et al.[28] found the same problems in a systematic review on cross-cultural adaptations and measurement property tests of a questionnaire designed to assess pain intensity (McGill Pain Questionnaire). Among the 44 different versions of the questionnaire for 26 different languages/cultures, most measurement properties were either not tested or were tested inadequately. The same was found in a systematic review of assessment tools designed for the evaluation of low back pain[29], for which most studies evaluated reliability and structural validity, but failed to test internal consistency, responsiveness, the ceiling effect and the floor effect.

Bot et al.[30], conducted a systematic review to analyze the measurement properties of questionnaires that assessed shoulder disorders and found different assessment methods for measurement properties, as well as flaws in assessments of the structural validity, internal consistency and reliability. In most of the instruments assessed, hypotheses related to the range and direction of the expected correlations with other instruments were not formulated. Factorial analysis was also not conducted. When it was used, it did not always confirm the dimensions that the questionnaire proposed to measure. Responsiveness was usually tested in samples of inadequate sizes. Most of the studies did not adequately describe the study method and/or data analysis.

It should be pointed out that other guidelines can be used for the evaluation of procedures for testing measurement properties that do not require all the criteria found on the COSMIN checklist. However, the decision was made to employ this checklist based on the fact that its quality criteria are the most updated and widely accepted in the literature[10-14,26].

Every effort was made in the systematic search of the electronic databases to identify studies on TMD assessment tools in Brazilian Portuguese. However, it is possible that unpublished data on the measurement properties of the assessment tools analyzed could be found in dissertations and theses. Such texts were not considered in the selection process, which may be interpreted as a possible limitation of the present review.

The measurement properties of the instruments included in this systematic review ranged in classification from good to weak, according to the criteria of the COSMIN checklist. These questionnaires are used in many Brazilian scientific epidemiological or clinical researches. Thus, care must be taken when interpreting the scores of questionnaires that have not had their measurement properties completely tested or were not tested in accordance with quality criteria.

Finally, we recommend that instrument researchers consider conducting full psychometric tests of their instruments using adequate sample sizes. We also recommend they consider scoring methods and quality criteria to provide scientifically robust instruments that are easy to administer.

References

1. Dworkin Sf, Hunggins Kh, LeResche L, Von Korff M, Howard J, Truelove E, et al. Epidemiology of signs and symptoms in temporomandibular disorders: clinical sings in cases and controls. J Am Dent Assoc. 1990 Mar;120(3):273-81.

2. Fonseca DM, Bonfate G, Valle AL, Freitas SFT. [Diagnosis of the craniomandibular disfunction through anamnesis]. RGO. 1994;42(1):23-8. Portuguese.

3. Helkimo M. Studies on function and dysfunction of the mastigatory system, II: índex for anamnestic and clinical disfunction and occlusal state. Sven Tandlak Tidskr. 1974 Mar;67(2):101-21.

4. Fricton JR, Schiffman EL. The craniomandibular index: validity. J Prosthet Dent. 1987 Aug;58(2):222-8.

5. Pehling J, Schiffman E, Look J, Shaefer J, Lenton P, Fricton J. Interexaminer reliability and clinical validity of the temporomandibular index: a new outcome measure for temporomandibular disorders. J Orofac Pain. 2002 Fall;16(4):296-304.

6. Truelove EI, Sommers Ee, Leresche L, Dworkin Sf, Von Korff M. Clinical diagnostic criteria for TMD: new classification permits multiple diagnoses. J Am Dent Assoc. 1992 Apr;123(4):47-54.

7. Dworkin SF, Leresche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord. 1992 Fall;6(4):301-55.

8. Keszei A, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. 2010 Apr;68(4):319-23. doi: 10.1016/j.jpsychores.2010.01.006.

9. Cano SJ, Hobart JC, Fitzpatrick R, Bhatia K, Thomp¬son AJ, Warner TT. Patient-based outcomes of cervical dystonia: a review of rating scales. Mov Disord. 2004 Sep; 19(9):1054-9.

10. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: COnsensus based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol. 2006 Jan;6:2.

11. Mokkink LB, Terwee Bb, Gibbons E, Stratford PW, Alonso J, Patrick DL, Knol DL, Bouter LM, De Vet HCW. Inter rater reliability of the COSMIN (COnsensus based Standards for the selection of health status Measurement Instruments) checklist. BMC Research Methodology. 2010;10:82.

12. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010 May;19(4):539-49. doi: 10.1007/s11136-010-9606-8..

13. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol. 2010 Mar 18;10:22. doi: 10.1186/1471-2288-10-22.

14. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. International consensus on taxonomy, terminology, and definitions of measurement properties for health related patient reported outcomes: results of the COSMIN study. J Clin Epidemiol. 2010 Jul;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006.

15. Lucena LBS, Kosminsky M, Costa LJ, Góes PSA. Validation of the Portuguese version of the RDC/TMD Axis II questionnaire. Braz Oral Res.2006 Oct-Dec;20(4):312-7.

16. Campos JADB, Carrascosa AC, Loffredo LCM, Faria JB. [Internal consistency and reproducibility of Portuguese version of research diagnostic criteria for temporomandibular disorders (RDC/TMD-Axis II)]. Rev Bras Fisioter 2007 Nov-Dec;11(6):451-9. Portuguese.

17. Cavalcanti RF, Studart LM, Kosminsky M, Goes PSA. Validation of the multimedia version of the RDC/TMD axis II questionnaire in Portuguese. J Appl Oral Sci. 2010 May-Jun;18(3):231-6.

18. Campos JADB, Carrascosa AC, Maroco J. Validity and reability of the Portuguese version of Mandibular Function Impairment Questionnaire. 2012 May;39(5):377-83. doi: 10.1111/j.1365-2842.2011.02276.x.

19. Campos JADB, Gonçalves DAG, Camparis CM, Speciali JG. Reliability of a questionnaire for diagnosing the severity of temporomandibular disorder. Rev Bras Fisioter. 2009;13(1):38-43.

20. Ferreira-Bacci AV. Mazzetto MO, Fukusima SS. [Adaptation of CR10 to Portuguese language for measuring pain in temporomandibular dysfunctions]. Rev Bras Odontol. 2009 Jan-Jun;66(1):16-21. Portuguese.

21. Campos JA, Carrascosa AC, Bonafé FS, Maroco J. Severity of temporomandibular disorders in womem: validity and reliability of the Fonseca Anamnestic Index. Braz Oral Res. 2014;28:16-21.

22. Pinelli C, Loffredo LCM. Reproducibility and validity of self-perceived oral health conditions. Clin Oral Investig. 2007 Dec;11(4):431-7.

23. Kallás MS, Crosato EM, Biazevic MGH, Mori M, Aggarwal VR. Translation and cross-cultural adaptation of Manchester Orofacial Pain Disability Scale. J Oral Maxillofac Res. 2013 Jan 1;3(4):e3. doi: 10.5037/jomr.2012.3403.

24. Manfredi APS, Da Silva AA, Vendite LL. [The sensibility appreciation of the questionnaire for selection of orofacial pain and temporomandibular disorders recommended by the American Academy of Orofacial Pain]. Rev Bras Otorrinolaringol. 2001 Nov-Dec;67(6):763-8. Portuguese.

25. Franco-Micheloni AL, Fernandes G, Gonçalves DAG, Camparis CM. Temporomandibular disorders among Brazilian adolescents: reliability and validity of a screening questionnaire. J Appl Oral Sci. 2014 Jul-Aug;22(4):314-22.

26. Terwee CB, Bot SD, De Boer MR, Van Der Windt DA, Knol Dl Dekker J, et al. Quality criteria were proposed for mensurement of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34-42.

27. Krebs DE. Declare your ICC type. Phys Ther. 1986 Sep;66(9):1431.

28. Costa LCM, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol. 2009 Sep;62(9):934-43. doi: 10.1016/j.jclinepi.2009.03.019.

29. Costa LO, Maher CG, Latimer J. Self-report outcome measures for low back pain: searching for international cross-cultural adaptations. Spine (Phila Pa 1976). 2007 Apr 20;32(9):1028-37.

30. Bot SD, Terwee CB, Van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis. 2004 Apr;63(4):335-41.