Puga, Vanessa O. O.; Lopes, Alexandre D.; Costa, Leonardo O. P.
Assessment of cross-cultural adaptations and measurement properties of self-report outcome measures relevant to shoulder disability in Portuguese: a systematic review
Associação Brasileira de Pesquisa e Pós-Graduação em Fisioterapia
São Carlos, Brasil

Available in: http://www.redalyc.org/articulo.oa?id=235023676002
Assessment of cross-cultural adaptations and measurement properties of self-report outcome measures relevant to shoulder disability in Portuguese: a systematic review

Avaliação das adaptações transculturais e propriedades de medida de questionários relacionados às disfunções do ombro em língua portuguesa: uma revisão sistemática

Vanessa O. O. Puga¹, Alexandre D. Lopes¹, Leonardo O. P. Costa¹,²

Abstract

Objectives: To evaluate the quality of the adaptation procedures as well as the clinimetric testing of the shoulder disability questionnaires available in Portuguese that has occurred for each adaptation. Methods: Systematic literature searches on MEDLINE, EMBASE, CINAHL, SCIELO and LILACS were performed to identify relevant studies. Data on the quality of the cross-cultural adaptation procedures and clinimetric testing were extracted. All studies were evaluated according to the current guidelines for cross-cultural adaptation and measurement properties. Results: Seven different questionnaires adapted into Brazilian-Portuguese (DASH, WORC, SPADI, PSS, ASORS, ASES and UCLA) were identified from eleven studies. Most of the studies performed the cross-cultural adaptation procedures following the recommendations from the guidelines. From a total of seven instruments, two were not tested for any measurement property (PSS and ASES) and two questionnaires (DASH and WORC) were evaluated for almost all of properties. None of the questionnaires were fully tested for their measurement properties. Conclusions: Although most of the shoulder disability questionnaires have been properly adapted into Brazilian-Portuguese, some of them were either inadequately tested or not tested at all. It is recommended that only tested instruments can be used in clinical practice, as well as in research.

Keywords: questionnaire; translation; validation studies; shoulder; physical therapy.

Resumo

Objetivos: Avaliar os procedimentos de tradução/adaptação cultural e das propriedades de medida de questionários que avaliam dor e disfunções no ombro, os quais já foram traduzidos/adaptados para a língua portuguesa. Métodos: Foram realizadas buscas sistematizadas nas bases de dados eletrônicas MEDLINE, EMBASE, CINAHL, SCIELO e LILACS para identificar os estudos relevantes. Foram extraídos os dados referentes à tradução e adaptação cultural, além dos dados das propriedades de medida de cada estudo. Todos os estudos foram analisados quanto à sua respectiva qualidade metodológica de acordo com as diretrizes para adaptação cultural e para as propriedades de medida. Resultados: Um total de 876 estudos foi identificado nas buscas, e, desses, apenas 11 foram considerados elegíveis, sendo que eles adaptaram e/ou testaram sete instrumentos diferentes (DASH, WORC, SPADI, PSS, ASORS, ASES e UCLA). A maioria deles cumpriu adequadamente as recomendações das diretrizes de adaptação transcultural. Dois dos sete questionários não foram testados para nenhuma propriedade de medida (PSS e ASES), e apenas dois questionários (WORC e DASH) foram testados para praticamente todas as propriedades de medida, porém nem todas foram testadas adequadamente. Nenhum questionário testou por completo todas as propriedades de medida. Conclusões: Os processos de tradução e adaptação transcultural foram realizados de maneira adequada para a maioria dos instrumentos, porém a maioria não teve suas propriedades de medida testadas adequadamente. Recomenda-se que somente instrumentos testados para suas respectivas propriedades de medida sejam utilizados na prática clínica assim como em pesquisas.

Palavras-chave: questionários; tradução (produto); estudos de validação; ombro; fisioterapia.

Received: 06/12/2011 – Revised: 09/10/2011 – Accepted: 09/26/2011

¹Masters Program in Physical Therapy, Universidade Cidade de São Paulo (UNICID), São Paulo, SP, Brazil
²Musculoskeletal Division, The George Institute for Global Health, Sydney, NSW, Australia
Correspondence to: Leonardo Oliveira Peña Costa, Rua Cesário Galeno, 48/475, Tatuape, CEP 03071-000, São Paulo, SP, Brasil, e-mail: lcosta@edu.unicid.br
Introduction

Shoulder pain is considered the third most common musculoskeletal condition. The prevalence of shoulder disorders ranges from 7 to 36% in the general population\textsuperscript{12}. Shoulder joint dysfunctions are often observed in workers and athletes who are exposed to repetitive movements and excessive strain, but some studies also report a high prevalence of this problem in elderly and sedentary people\textsuperscript{2-4}.

Assessment methods for musculoskeletal injuries have been modified in recent years\textsuperscript{5}. Instead of a purely physical examination, including tests of muscular strength, joint mobility and imaging exams, self-report questionnaires or scales were incorporated\textsuperscript{6,7}. Questionnaires are widely used to collect information related to important clinical outcomes, such as pain intensity, quality of life, satisfaction with treatment and disability\textsuperscript{8,9}. In addition to the use of questionnaires in clinical practice, they are also widely used in research\textsuperscript{10,11}.

There are several instruments for assessing patients with shoulder dysfunctions and detecting changes in their clinical condition over time, but most of them were developed in English\textsuperscript{5,12}. In order to be used in Brazil, they must be translated, cross-culturally adapted and tested for their measurement properties such as the internal consistency, reproducibility, validity and responsiveness\textsuperscript{13-15}.

The translation and cross-cultural adaptation process is necessary to verify the equivalence with the original version and to resolve cultural, linguistic and health perception differences between different countries and cultures\textsuperscript{15}. Even if an instrument’s measurement properties have already been tested in its original version, it is necessary to retest them after translation, since cultural differences may affect the results. Testing is also necessary to determine if the adapted questionnaire retains the measurement properties of the original version\textsuperscript{14}. Guidelines for evaluating the appropriateness translation and cross-cultural adaptation\textsuperscript{13} and clinimetric testing\textsuperscript{13-15} have been developed in order to help researchers to conduct optimal studies on these topics.

Some questionnaires specifically designed to assess shoulder dysfunctions have already been translated and cross-culturally adapted into Portuguese and have had some of their measurement properties tested; however, a synthesis of the information contained in these questionnaires is still not available in the literature. Therefore the objectives of this study were to describe and to assess the translation and to cross-cultural adaptation procedures of these instruments and to describe and to assess the measurement properties tested in each of the studies.

Methods

Study selection

In order to identify the instruments available in Portuguese for assessing shoulder pain and disability, systematic searches were performed in five electronic databases (MEDLINE through OVID, EMBASE, CINAHL through EBSCO, SCIELO and LILACS). The search terms and Boolean operators (AND, OR or NOT) used in the databases MEDLINE, EMBASE, CINAHL and SCIELO were: (shoulder OR impingement OR rotator cuff OR arm OR upper limb OR instability OR upper extremity) AND (questionnaire OR index OR scale OR score OR assessment OR evaluation OR self report OR inventory) AND (Brazil OR Brasil OR Portuguese OR Brazilian Portuguese OR Brazilian). The search terms (in Portuguese) used in the LILACS database were: (ombro OR manguito rotador OR membro superior OR instabilidade) AND (questionário OR escala OR índice OR instrumento OR escore OR avaliação) AND (Brasil OR português OR português brasileiro). The searches were not limited by language or publication date. The final search was performed in April 2011.

Inclusion criteria

Only instruments developed for the assessment of shoulder joint dysfunctions that had been translated and/or cross-culturally adapted into Portuguese were included in the study, regardless of whether they were combined with the assessment of other upper limb joints. Self-report instruments and evaluator-dependent instruments using only objective measurements, such as strength or range of motion, were also included in this review. Only full-text papers were included; theses/dissertations, abstracts from conferences and books were excluded from this systematic review.

Data extraction and assessment of methodological quality of eligible studies

Data regarding the translation and cross-cultural adaptation were extracted in order to assess the design of these procedures. We also extracted data relating to the measurement properties of each study, including the sample size, reproducibility (reliability and agreement), responsiveness (and time interval between measures), internal consistency and construct validity (raw data can be requested from the corresponding author via email).

Subsequently, the translation and cross-cultural adaptation methods of each study were classified according to the
Guidelines for the Process of Cross-cultural Adaptation of Self-report Measures\textsuperscript{13}. The translation and adaptation process includes a initial translation, a synthesis of translations, back-translation, reviews by an Expert Committee and the pre-test version of the instrument. The quality of each step was classified as either positive (+) when the procedure was performed in accordance with the quality criteria, doubtful (?) when the description of method was unclear, negative (-) when the procedure was performed correctly but with an insufficient quantity of translators and/or back-translators or, finally, zero (0) when there was not enough information to evaluate each step (Table 1).

The measurement properties were classified according to the Quality Criteria for Measurement Properties of Health Status Questionnaires\textsuperscript{14} and the evaluation was restricted only to those items relevant to the instruments evaluated. Other original items of the Quality Criteria such as content (or face) validity and interpretability are relevant for the development of the original questionnaire. Similarly, the item criterion validity should be considered only when there a gold standard available for comparison, which is not the case for shoulder assessment instruments, so these three items were not included in our review.

The items assessed in this review were construct validity, internal consistency, reproducibility (agreement and reliability), responsiveness and ceiling and floor effects. Similarly to the cross-cultural adaptation criteria, the quality of each of the measurement properties was also classified as positive (+) when the procedures for each stage were performed in accordance with the criteria, doubtful (?) when the methods or the design of the study were unclear, negative (-) when the data for each clinimetric property included greater or smaller values than those defined by the criteria in spite of appropriate approach or methods, or zero (0) when there was no information to qualify each measurement property (Table 2). Data extraction and the assessments were carried out by one rater and then checked by an independent reviewer, who reviewed all data. There was no disagreements between the rater and the independent reviewer who met and discussed the data.

### Results

A total of 876 studies were retrieved from the searches but only 11 were considered eligible for data analysis (Figure 1). From these, seven different instruments that had been translated and cross-culturally adapted into Portuguese were identified: DASH (Disabilities of the Arm, Shoulder and Hand)\textsuperscript{16}, SPADI (Shoulder Pain and Disability Index)\textsuperscript{17}, WORC (Western Ontario Rotator Cuff Index)\textsuperscript{18}, ASES (American Shoulder and Elbow Surgeons Questionnaire)\textsuperscript{19}, PSS (Penn Shoulder Score)\textsuperscript{20}, ASORS (Athletic Shoulder Outcome Rating Scale)\textsuperscript{21} and Modified-UCLA (Modified-University of California at Los Angeles Shoulder Rating Scale)\textsuperscript{22}. Of the seven instruments translated and cross-culturally adapted into Portuguese, five had been tested for at least one measurement property: DASH\textsuperscript{16,22}, WORC\textsuperscript{18,22-24}, SPADI\textsuperscript{27},

---

### Table 1. Guidelines for the process of cross-cultural adaptation of self-report measures\textsuperscript{13} (adapted from Costa et al.\textsuperscript{23}).

<table>
<thead>
<tr>
<th>Steps</th>
<th>Description</th>
<th>Rating Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>Two (or more) translators should independently translate the original</td>
<td>+ Translation performed by at least two independent translators</td>
</tr>
<tr>
<td></td>
<td>questionnaire. The translators should preferably be native speakers to the</td>
<td>- Doubtful translation procedure</td>
</tr>
<tr>
<td></td>
<td>target language.</td>
<td>0 No information about translation</td>
</tr>
<tr>
<td>Synthesis</td>
<td>The translators should synthesize the multiple translations to produce a</td>
<td>+ Performed synthesis</td>
</tr>
<tr>
<td></td>
<td>consensus of the translations.</td>
<td>- Doubtful design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 No information about synthesis OR translation performed by only one translator</td>
</tr>
<tr>
<td>Back translation</td>
<td>Translators, blinded to the original questionnaire, should translate the</td>
<td>+ Back translation performed by at least two independent translators</td>
</tr>
<tr>
<td></td>
<td>consensus translation back into the original language.</td>
<td>- Doubtful back translation procedure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 No information about back translation</td>
</tr>
<tr>
<td>Expert committee</td>
<td>The expert committee should consolidate all the versions of the questionnaire and develop what would be considered the prefinal version of the questionnaire for testing.</td>
<td>+ Clearly reported the existence of an expert committee</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Doubtful design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 No information about expert committee</td>
</tr>
<tr>
<td>Pretesting</td>
<td>The prefinal questionnaire undergoes pilot testing with members of the target population.</td>
<td>+ Performed pretesting</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Doubtful design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 No information about pretesting</td>
</tr>
</tbody>
</table>

+=positive rating; -=negative rating; 0=no information available; ?=unclear.
Table 2. Quality criteria for measurement properties of health status questionnaires\(^4\) (adapted from Costa et al.\(^3\)).

<table>
<thead>
<tr>
<th>Property</th>
<th>Definition</th>
<th>Quality Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal consistency</td>
<td>Internal consistency is a measure of the homogeneity of a (sub)scale. It indicates the extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct. Factor analysis should be applied to determine the dimensionality of the item—this is, to determine whether or not they formed only one overall dimension or more than one.</td>
<td>+ Factor analyses performed on adequate sample size (7 x # items and ≥100) AND Cronbach’s alpha(s) calculated per dimension AND Cronbach’s alpha(s) between 0.70 and 0.95; - No factor analysis OR doubtful design or method; - Cronbach’s alpha(s) &lt;0.70 or &gt;0.95, despite adequate design and method; 0 No information found on internal consistency.</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Content validity examines the extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured.</td>
<td>+ Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses; ? Doubtful design or method (e.g., no hypotheses); - Less than 75% of hypotheses were confirmed, despite adequate design and methods; 0 No information found on construct validity.</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>The degree to which repeated measurements in stable persons (test retest) provide similar answers.</td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td>The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)</td>
<td>+ ICC or Kappa ≥0.70; ? Doubtful design or method (e.g., time interval not mentioned); - ICC or Kappa &lt;0.70, despite adequate design and method; 0 No information found on reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>The extent to which the scores on repeated measures are close to each other (absolute measurement error)</td>
<td>+ MIC &lt;SDC or MIC outside the LOA or convincing arguments that agreement is acceptable; ? Doubtful design or method or (MIC not defined AND no convincing arguments that agreement is acceptable); - MIC =SDC or MIC equals or inside LOA, despite adequate design and method; 0 No information found on agreement.</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>The ability of a questionnaire to detect clinically important change over time in the concept being measured. A predefine hypotheses about the relation of change in the instrument to corresponding changes in reference measures should be postulated.</td>
<td>+ Smallest detectable change individual or Smallest detectable change group &lt; Minimal important change OR Minimal important change outside the limits of agreement OR Responsiveness ratio &gt;1.96 OR Area under the curve ≥0.70; ? Doubtful design or method OR sample size &lt; 50 OR methodological flaws; - Smallest detectable change individual or Smallest detectable change group ≥ Minimal important change OR Minimal important change equals or inside limits of agreement OR Responsiveness ratio ≤1.96 OR Area under the curve &lt;0.70, despite adequate design and methods; 0 No information found on agreement.</td>
</tr>
<tr>
<td>Floor and ceiling effects</td>
<td>The number of respondents who achieved the lowest or highest possible score</td>
<td>+ ≤15% of the respondents achieved the highest or lowest possible scores; ? Doubtful design or method OR sample size &lt;50 OR methodological flaws; - &gt;15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods; 0 No information found on interpretation.</td>
</tr>
</tbody>
</table>

+=positive rating; ?=doubtful design or method; -=negative rating; 0=no information available. Doubtful design or method=lacking of a clear description of the design or methods of the study, sample smaller than 50 subjects, or any important methodological weakness in the design or execution of the study. MIC=minimal important changes; SDC=smallest detectable change; LOA=limits of agreement; ICC=interclass correlation coefficient; SD=standard deviation.

ASORS\(^25\) and Modified-UCLA\(^24\). Finally, the PSS\(^20\) and the ASES\(^19\) had not been tested for any measurement property.

Table 3 presents the ratings of the translations and cross-cultural adaptations according to the Guidelines for the Process of Cross-cultural Adaptation of Self-report Measures\(^13\). For the seven instruments found in the 11 included studies, the steps regarding translation, synthesis, analysis by a committee of experts and the pre-tests had all been correctly performed. The back-translation step had not been adequately performed for the WORC and ASORS as it was performed by a single
Table 4 presents the ratings of all evaluated measurement properties according to the Quality Criteria for Measurement Properties of Health Status Questionnaires\textsuperscript{14}. Reliability was the most frequently tested measurement property, included in four of the five studies that tested measurement properties. These four studies performed the test appropriately and obtained Intraclass Correlation Coefficient (ICC) estimates ≥0.70\textsuperscript{16,17,23,25}. On the other hand, agreement was tested only for the WORC in a study that also tested its reliability, and it presented appropriate estimates\textsuperscript{23}. Internal consistency was tested for DASH\textsuperscript{26}, SPADI\textsuperscript{17} and WORC\textsuperscript{23}; the first two presented acceptable levels (Cronbach’s alpha ranging between 0.70 and 0.95), but the WORC’s design was questionable, since the internal consistency was tested by intra- and interrater reliability rather than the criteria suggested by the guidelines\textsuperscript{14}. Factorial analysis was performed only for the DASH. Construct validity was not properly tested in any of the instruments found in this review (DASH\textsuperscript{16,22}, WORC\textsuperscript{23,24} and ASORS\textsuperscript{25}) due to the fact that the hypothesis regarding the correlations of the instruments/scales had not been formulated a priori. Responsiveness was tested in one study that involved DASH, WORC, and UCLA\textsuperscript{24}, but it could not be considered appropriate due to a small sample size (i.e. lower than 50 participants). Ceiling and floor effects were not tested in any of the instruments identified in our systematic review.

**Table 3.** Cross-cultural adaptations of the shoulder questionnaires adapted into Brazilian-Portuguese that used the translation-based approach related to the Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Translation</th>
<th>Synthesis</th>
<th>Back translation</th>
<th>Expert committee review</th>
<th>Pretesting</th>
</tr>
</thead>
<tbody>
<tr>
<td>DASH\textsuperscript{16}</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>UCLA\textsuperscript{21}</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>WORC\textsuperscript{18}</td>
<td>+</td>
<td>+</td>
<td>_*</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>WORC\textsuperscript{23}</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DASH\textsuperscript{26}</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>WORC e DASH\textsuperscript{24}</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>WORC, DASH e UCLA\textsuperscript{24}</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ASES\textsuperscript{19}</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>PSS\textsuperscript{20}</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SPADI\textsuperscript{17}</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ASORS\textsuperscript{25}</td>
<td>+</td>
<td>+</td>
<td>_*</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

DASH—Disabilities of the Arm, Shoulder and Hand; UCLA—University of California Los Angeles Shoulder Rating Scale; WORC—Western Ontario Rotator Cuff Index; ASES—American Shoulder and Elbow Surgeons Questionnaire; PSS—Penn Shoulder Score; ASORS—Athletic Shoulder Outcome Rating Scale; SPADI—Shoulder, Pain and Disability Index. N/A—Not applicable — The cross cultural adaptations was not performed, only the clinimetric tests. The questionnaires used in these studies have been previously translated in other studies. * Back translation performed by only one translator.
Almost all of the measurement properties were tested for the DASH and WORC; only ceiling and floor effects were not tested for the WORC22-24, whereas both ceiling and floor effects and agreement were not tested for the DASH20,23,24,26.

Two measurement properties were tested for both SPADI and ASORS: reliability and internal consistency were tested properly for SPADI, and reliability was tested adequately in ASORS25, but construct validity was inadequately conducted25. Only responsiveness was evaluated for UCLA, but it was rated as doubtful24.

### Discussion

The objectives of this study were to describe and assess the translation and cross-cultural adaptation procedures of these instruments and to describe and assess the measurement properties tested in each of the studies. Seven different instruments (from 11 studies) were retrieved and at least one measurement property was tested in five of them. The translation and adaptation processes were assessed in accordance with the Guidelines for the Process of Cross-cultural Adaptation of Self-report Measures13; the measurement properties were assessed according to the Quality Criteria for Measurement Properties of Health Status Questionnaires14.

The guidelines for cross-cultural adaptations13 was followed for DASH16, SPADI17, UCLA23, PSS20 and ASES19 with all steps having been adequately performed. The specific protocol for linguistic equivalence suggested by the authors of the original WORC18 was followed, according to the criteria of the MAPI Research Institute27 and therefore back-translation was performed by a single translator. The authors of the Portuguese WORC study reported that the guidelines proposed by Guillemim, Bombardier and Beaton28 and Beaton et al.13 involved difficult processes and due to the population tested, the complexity of following each of the steps, the long duration to perform all steps and high costs they followed a synthesized protocol, pointing out that all versions of WORC under development in other languages followed the same procedures. The guidelines proposed by Guillemim, Bombardier and Beaton28 were followed for the ASORS25, but the guideline-recommended back-translation step was not fully performed because only one translator was involved. All other steps were carried out adequately in both questionnaires.

None of the questionnaires have tested all measurement properties. In fact, no measurement properties were evaluated in two of the seven instruments (PSS and ASES). Testing was conducted for virtually all measurement properties for the WORC and DASH, however these tests did not always occurred in accordance with the guidelines.

Reliability was the most frequently tested property. The tests were applied correctly in all studies16,17,23,25, including adequate sample size and adequate measurement intervals14, however none of the studies mentioned which type of ICC had been used while testing reliability. Only the studies that tested the reliability of SPADI17 and ASORS25 described their 95%
confidenc intervals. It is extremely important to report the
type of ICC used in different tests, since different ICCs might
lead to completely distinct results, under- or overestimating the 
estimates of reliability29.

Agreement is an important measurement property and re-
fects the degree to which repeated measures applied to stable
patients provide similar answers14. Unfortunately our review
identified only one study23 that tested agreement. Agreement
is more easily interpreted clinically than reliability due to the
fact that it is expressed on the instrument’s units of measure-
ment. Reliability, on the other hand, presents its indexes (ICC
or Kappa) expressed on a scale ranging from 0 to 1, which, in
many cases, require a more difficult interpretation. In order
to fully test reproducibility it would be ideal to test both the
reliability (relative error of measurement) and the agreement
(absolute error of the measure), unfortunately most studies
tested reliability only.

Construct validity was tested in three studies (DASH16,22,
WORC23 and ASORS23). All of them used Pearson Correlation
tests, which involve correlating a questionnaire with other
similar instruments. However, hypotheses must be formulated
a priori and they must specify both the magnitude and direc-
tion of the expected correlation14. Such formulated hypotheses
were found in none of the included studies that tested con-
struct validity14,22,23,25. A specific a priori hypothesis is necessary
because it would be easier to develop an alternative explana-
tion for low correlations than admit that the questionnaire’s
levels of construct validity is compromised14.

Internal consistency was tested for DASH16, WORC23 and
SPADI17. Only the study that assessed the internal consist-
cency of DASH26 used Cronbach’s alpha in combination with
factorial analysis, identifying three different factors. This
analysis is important because such a procedure can identify
how many factors are present in a questionnaire and if there
is more than one factor Cronbach’s alpha should be calcu-
lated for each factor separately14. The studies that evaluated
WORC23 and SPADI17 used Cronbach’s alpha but not facto-
rial analysis.

Responsiveness was evaluated only in one study that
tested DASH, WORC and UCLA24 measurement properties.
This study used an appropriate between-measure interval
(i.e. three months) but recruited a small sample size
(30 patients). Responsiveness represents the ability of a
questionnaire to detect clinical changes over time and
can be measured by internal (measured by effect sizes) or
external responsiveness (measured by correlation tests
and/or the construction of receiver operator characteristics
curves)24. No study in this systematic review fully tested
responsiveness. Moreover, no study tested ceiling and
floor effects and, consequently, it is not known whether
the evaluated instruments would fail to detect patient
improvement or deterioration14,30.

Other systematic reviews of measurement instruments
confirm our findings that there is a clear need to complete the
evaluation of the instrument’s measurement properties
so that the best choice of questionnaire for specific situations
can be made11,32. In a systematic review of English language
shoulder disability questionnaires9, different methods of as-
sessing measurement properties were found, in addition to
flaws in the construct validity, internal consistency and re-
ponsiveness assessments.

In most evaluated instruments, hypotheses about the
expected magnitude and direction of correlations with
other instruments had not been formulated. Factorial
analyses were not generally used but, even when present
the dimensions that the questionnaire intended to measure
were lacking in some cases. Responsiveness was frequently
tested, although with an inadequate sample size (n<43)
and finally most studies did not adequately describe the
study methods and/or data analysis.

The same types of problems were also found in a system-
atic review of cross-cultural adaptations and measurement
properties of the McGill Pain Questionnaire, which is an
instrument for assessing the quality and intensity of pain11.
Among the 44 different versions of the questionnaire, rep-
resenting 26 different languages/cultures, it was frequently
observed that the tests had been not performed or were
conducted after publication of the
Guidelines for the Process
of Cross-cultural Adaptation of Self-report Measures13, and only
one study, which tested DASH measurement properties (re-
liability and construct validity)26 was developed prior to the
publication of the Quality Criteria for Measurement Properties
of Health Status Questionnaires14.

We have used all efforts in our searches in order to
identify and retrieve all relevant questionnaires that were
adapted into Brazilian-Portuguese. Even though the most

Systematic review of shoulder questionnaires

quality criteria due to fact that these guidelines are the most
updated and widely accepted in the literature. All of the trans-
lation and adaptation studies examined in this review were
conducted after publication of the
Guidelines for the Process
of Cross-cultural Adaptation of Self-report Measures13, and only
one study, which tested DASH measurement properties (re-
liability and construct validity)26 was developed prior to the
publication of the Quality Criteria for Measurement Properties
of Health Status Questionnaires14.

We have used all efforts in our searches in order to
identify and retrieve all relevant questionnaires that were
adapted into Brazilian-Portuguese. Even though the most
References


