UCSF
UC San Francisco Previously Published Works

Title

Systematic Review and Appraisal of the Cross-Cultural Validity of Functional Status Assessment Measures in Rheumatoid Arthritis

Permalink

https://escholarship.org/uc/item/0p67x17f

Journal

Arthritis Care & Research, 72(6)

ISSN

2151-464X

Authors

Kulhawy-Wibe, Stephanie C
Zell, JoAnn
Michaud, Kaleb
et al.

Publication Date

2020-06-01

DOI

10.1002/acr.23904

Peer reviewed

AMERICAN COLLEGE
of RHEUMATOLOGY
Empowering Rheumatology Professionals

Systematic Review and Appraisal of the Cross-Cultural Validity of Functional Status Assessment Measures in Rheumatoid Arthritis

Stephanie C. Kulhawy-Wibe,[1] JoAnn Zell,[2] Kaleb Michaud,[3] [ID] Jinoos Yazdany,[4] Aileen M. Davis,[5] Linda Ehrlich-Jones,[6] J. Carter Thorne,[7] Donna Everix,[8] Laura C. Cappelli,[9] Lisa G. Suter,[10] Alex Limanni,[11] and Claire E. H. Barber[12] [ID]

Objective. We conducted a systematic review and appraisal of the cross-cultural adaptation and cross-cultural validity of the Health Assessment Questionnaire (HAQ) and its derivatives, and of the more recent Patient-Reported Outcomes Measurement Information System (PROMIS) functional status assessment measures (FSAMs) in rheumatoid arthritis.

Methods. Four electronic medical databases were searched from inception until April 4, 2018 according to the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) group search strategy. Included studies were evaluated using the COSMIN tool for cross-cultural validity and were scored as excellent, good, fair, or poor.

Results. Of 58 articles identified by our search strategy and 3 by manual search, 39 were included: 29 described the translation, cultural adaptation, or cross-cultural validity of the HAQ disability index, 8 other HAQ derivatives, and 2 PROMIS measures, representing 22 languages. Of the 39 articles reviewed, 3 examined the cross-cultural validity of translated versions. These studies were rated as follows: 2 as excellent, 3 good, 13 fair, and 21 poor. Two studies examining cross-cultural validity noted differential item functioning (DIF) between Dutch and US populations for the HAQ-II and PROMIS measures, and a third study found DIF between Turkish and UK populations on the HAQ, indicating cultural differences in questionnaire response.

Conclusion. This review highlights a paucity of data on the cross-cultural validity of FSAMs and the mostly poor- or fair-quality methods by which they were translated and adapted, which needs to be considered when using these measures for multinational clinical trials and for day-to-day use in clinical practice.

INTRODUCTION

Functional status assessment measures (FSAMs) are important outcome measures in rheumatoid arthritis (RA), because poor function is a predictor for mortality, is associated with lower quality of life and with work disability (1–4), and is an important outcome in clinical treatment trials. The Health Assessment Questionnaire disability index (HAQ DI) (5) and its derivatives (6–8) are standardized and validated FSAMs commonly used in RA. More recently, the Patient-Reported Outcomes Measurement Information System (PROMIS) has been developed and includes FSAMs (9). The American College of Rheumatology (10) recommends repeated assessment of FSAMs in clinical practice to track patient outcomes and to guide shared decision-making, and FSAMs are used worldwide for this purpose.

The HAQ DI, HAQ-derived instruments, and PROMIS measures were developed and validated in English but have been translated and culturally adapted for use around the world (11).

[1]Stephanie C. Kulhawy-Wibe, MD, MSc: Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [2]JoAnn Zell, MD: Denver Health and University of Colorado, Denver; [3]Kaleb Michaud, PhD: University of Nebraska Medical Center, Omaha, and FORWARD, The National Databank for Rheumatic Diseases, Wichita, Kansas; [4]Jinoos Yazdany, MD, MPH: University of California, San Francisco; [5]Aileen M. Davis, PhD: Krembil Research Institute, University Health Network, and Institute of Health Policy, Management and Evaluation, Dalla Lana, School of Public Health, University of Toronto, Toronto, Ontario, Canada; [6]Linda Ehrlich-Jones, PhD, RN: Shirley Ryan AbilityLab, Chicago, Illinois; [7]J. Carter Thorne, MD, FRCPC: University of Toronto, Toronto, Ontario, Canada; [8]Donna Everix, MPA, BS, PT: On My Care Home Health, Fremont, California;

[9]Laura C. Cappelli, MD: John Hopkins University, Baltimore, Maryland; [10]Lisa G. Suter, MD: Yale University, New Haven, Connecticut; [11]Alex Limanni, MD: Arthritis Centers of Texas, Dallas; [12]Claire E. H. Barber, MD, PhD, FRCPC: Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada, and Arthritis Research Canada.

No potential conflicts of interest relevant to this article were reported.

Address correspondence to Claire E. H. Barber, MD, PhD, FRCPC, University of Calgary, Faculty of Medicine, Room 451A HMRB, 3330 Hospital Drive NW, Calgary, Alberta T2N 4N1, Canada. E-mail: cehbarbe@ucalgary.ca.

Submitted for publication March 11, 2019; accepted in revised form April 9, 2019.

SIGNIFICANCE & INNOVATIONS

- There are various Health Assessment Questionnaire and Patient-Reported Outcomes Measurement Information System functional status assessment measures (FSAMs) that have been translated and culturally adapted for use around the world, although there is little examination of the cross-cultural validity of these measures.
- This is the first study to conduct a critical appraisal of the translation and cultural adaptation of these measures using a standardized assessment checklist developed by the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) group, and a majority of studies were found to be of poor or fair methodologic quality.
- Given the widespread use of these measures, future studies should be conducted to ascertain whether the cultural validity of FSAMs varies by disease subtype or whether further specific studies are required in rheumatoid arthritis populations.

Guidelines exist to support this multistep process that have themselves been evolving over the last few decades (12–18). The steps for translating, culturally adapting, and assessing the cross-cultural validity of measures can be evaluated using a standardized assessment checklist (19) developed by the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) initiative (20). COSMIN is an international initiative that aims to improve the science of outcome measurement through the development of tools to evaluate the methodologic quality of studies on outcome measurement properties, guidelines for conducting systematic reviews in this area, and the development and selection of outcome measures (20).

While rigorous methods for translation and cultural adaptation of patient-reported measures are important, further steps are necessary to assess the cross-cultural validity of these measures, to ensure that the latent trait of the measure (in this case functional status) is being measured in the same way across cultures (21). Differential item functioning (DIF) is a quantitative method to assess for the systematic differences in responses to questionnaire items based on different population characteristics (i.e., in this case cultural characteristics) that are otherwise comparable with respect to the latent trait being measured (21). There are a number of statistical methods for assessing DIF, including but not limited to logistic regression and item response theory (IRT). IRT is a paradigm for developing, evaluating, and scoring measures based on a number of mathematical models. Central to IRT is modeling the relationship between a patient's response to a questionnaire item and the underlying trait being measured (22). The objective of this current study was to conduct a systematic review of the cross-cultural validity of FSAMs for RA, including HAQ DI, HAQ-derived measures, and PROMIS measures, and to assess their methodologic quality using the COSMIN cross-cultural validity checklist.

MATERIALS AND METHODS

This systematic review is part of a larger ongoing project designed to evaluate and generate recommendations for the use of RA FSAMs in routine clinical practice. This current study exclusively examines the cross-cultural validity of FSAMs in RA.

Search strategy. Four electronic medical databases, Medline, Embase, Cochrane Library, and the Cumulative Index to Nursing and Allied Health Literature, were searched from inception to March 16, 2017. This search was updated on April 4, 2018 to ensure the review included all current articles. The Preferred Reporting Items for Systematic Reviews and Meta-Analysis checklist was used to guide our systematic review (23). Search terms were chosen according to a published strategy for finding studies on measurement properties of measurement instruments by Terwee et al (24) from the COSMIN group (20). This strategy was initially developed in Medline using medical subject heading terms and keywords in 3 themes: construct search (for assessment of functional status), population search (rheumatoid arthritis), and instrument search (including terms for instruments of interest, e.g., questionnaires) that were then adapted accordingly for the other databases. These 3 themes were then combined using the Boolean search operator AND. A title and abstract screening was conducted independently and in duplicate (JZ and CEHB) for articles that pertained to cross-cultural validity. A full-text review by 2 independent reviewers (SCK-W and CEHB) was used to determine eligible studies for inclusion. Any disagreement between reviewers was resolved by discussion. Next, the reference lists of all included articles were manually searched to identify any additional relevant studies. Reference lists for published reviews on the topic and any included articles were also hand searched for additional relevant publications. Finally, international online databases for patient-reported outcome measures were searched for additional translated measures or any further potentially relevant studies. This search included databases such as the European League Against Rheumatism Outcome Measures Library (oml.eular.org), the Health Outcomes Group (healthoutcomesgroup.com), and Mapi Research Trust (mapi-trust.org).

Eligibility criteria and article selection. Studies that described the translation or cross-cultural validation of the following patient-reported FSAMs in RA were included: HAQ DI, HAQ-derived instruments, or PROMIS FSAMs. The following exclusion criteria were applied: non-English publications, studies of translated FSAMs in non-RA populations, and studies using a previously translated FSAM to validate another translated measure (i.e., the original measure was not used as the gold standard).

Data collection and assessment of methodologic quality using COSMIN. Two reviewers, SCK-W and CEHB, independently conducted data collection and assessment of the methodologic quality of the included studies, and any disagreements were resolved by consensus. Study characteristics were collected, including original and translated language, country where the study took place, the specific FSAM studied, and the methods for cross-cultural translation and for determining cross-cultural validity, where appropriate.

The methodologic quality of the included studies was rated using the COSMIN checklist for cross-cultural validity (19,25). Briefly, COSMIN is a standardized tool for assessing the methodologic quality of studies on all of the following measurement properties: internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing, cross-cultural validity, responsiveness, and interpretability. COSMIN reporting also includes standardized collection of items relating to the interpretability of the measurement property (including percentage of missing items, handling of missing items, and adequate sample size) and the generalizability of the study (including population characteristics and study setting).

The checklists are meant to be used in a modular fashion, and this study uses only 1 of these checklists: cross-cultural validity. For this tool, a checklist of 15 items was completed, with each item rated on a 4-point scale (poor, fair, good, or excellent) based on predefined criteria (25) (see Supplementary Appendix A, available on the *Arthritis Care & Research* web site at http://online library.wiley.com/doi/10.1002/acr.23904/abstract). If a reviewed study reported on the other aforementioned psychometric properties, in addition to cross-cultural validity, it was not evaluated as part of this review, because it was considered outside the scope of evaluation.

The COSMIN checklist for cross-cultural validity (see Supplementary Appendix A, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.23904/

abstract) (19,25) is based on published guidelines for the multistep process that begins with translation, followed by cultural adaptation, and finally by assessing cross-cultural validity of instruments for use in other languages and cultures (12–16,18). Direct linguistic translation ideally includes both forward and backward translation by at least 2 independent translators, followed by an independent review panel to rectify issues with these translations. The next key element in developing a successful translation is conducting a pilot study or pretesting. Pretesting is essential to check face validity by assessing interpretability and comprehension as well as cultural relevance of the translation. Pretesting is best accomplished through in-person cognitive interviews, because this technique allows the most complete assessment of the question and answer process (26). At the end of this process, the translated and culturally adapted measure must then be assessed for cross-cultural validity, commonly measured by DIF, a formal test of invariance across cultures.

RESULTS

The original search yielded 11,835 articles, which all underwent title and abstract screening. Of these, 58 were eligible for full-text review. The repeated search yielded 1,082 articles, none of which were eligible for full-text review. Of these 58 articles, 36 were included, and an additional 3 articles were identified during hand searching for a total of 39 articles. The results of the search are shown in Figure 1.

Table 1 shows the characteristics of the included studies. Of the 39 studies, 29 described the translation, cultural adaptation, or cross-cultural validity of the HAQ DI, 8 other studies described HAQ derivatives, and 2 studies described PROMIS measures. In total, these represented 22 different languages. Nearly all of the articles (90%, n = 35) described translation, most (80%, n = 31) described cultural adaptation, and only very few (8%, n = 3) examined the cross-cultural validity of translated versions. There was
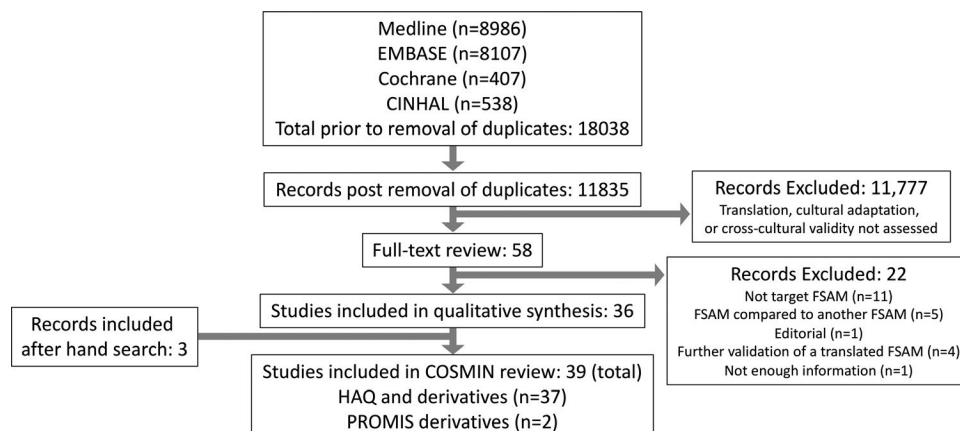

Figure 1. Flow diagram depicting manuscript selection for systematic review of cross-cultural validity of functional status assessment measures (FSAMs). CINAHL = Cumulative Index to Nursing and Allied Health Literature; COSMIN = Consensus-Based Standards for the Selection of Health Measurement Instruments; HAQ = Health Assessment Questionnaire; PROMIS = Patient-Reported Outcomes Measurement Information System.

Table 1. Characteristics of included studies examining the translation, cultural adaptation, or cross-cultural validity of common patient-reported functional status measures in rheumatoid arthritis*

Instrument	Studies, no.	Languages (number of studies)†	COSMIN rating	Translation	Cultural adaptation	Cross-cultural validity
HAQ DI	29	20 languages: Arabic (3), Bengali, British English, Chinese, Danish, Dutch, Estonian, French, Italian, Japanese, Korean, Malay, Marathi, Persian (3) Portuguese, Spanish (4), Swedish, Swiss-German, Thai (2), and Turkish (2)	Poor 17 (59), Fair 9 (31), Good 2 (7), Excellent 1 (3)	27 (93)	26 (90)	1 (3)
MDHAQ, MHAQ, HAQ-II	8	8 languages: Arabic, Chinese, Dutch, Finnish, Hindi, Korean, Spanish, and Swedish	Poor 3 (38), Fair 4 (50), Good 1 (13)	7 (88)	4 (50)	1 (13)
PROMIS	2	1 language: Dutch (2)	Poor 1 (50), Excellent 1 (50)	1 (50)	1 (50)	1 (50)

* Values are the number (%) unless indicated otherwise. COSMIN = Consensus-Based Standards for the Selection of Health Measurement Instruments; HAQ DI = Health Assessment Questionnaire disability index; MDHAQ = Multidimensional Health Assessment Questionnaire; MHAQ = modified HAQ; PROMIS = Patient-Reported Outcomes Measurement Information System.
† The number of studies per language is given if >1, including dialects.

generally poor concordance with the proposed guidelines for cross-cultural adaptations of the measures. However, a trend to improved study quality in the domain of cross-cultural adaptation was seen over time, with fewer studies published after 2009 rated as poor (2 of 9 poor quality studies in the time period 2009–2017 versus 18 of 26 in the time period between 1986–2008) (see Supplementary Appendices B and C, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.23904/abstract). Overall, only 2 studies (5%) were rated as excellent, 3 studies (7%) were rated as good, 13 (33%) were rated as fair, and 21 (54%) were rated as poor (see full ratings in Supplementary Appendices B and C, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.23904/abstract). Lack of pretesting was the main methodologic issue leading to a poor COSMIN rating in 43% of the reviewed articles (n = 16 of 37 eligible studies). Pretest-

ing the adapted measure refers to a methodologic step where the interpretation and cultural relevance of the items and comprehension are checked often through cognitive interviews. The characteristics of the population in which the pretesting is done should also be described. Another common issue was translation. The expertise of the people involved in translation was not described in 49% of the 35 studies reporting on translation (n = 17), resulting in a fair rating (see Supplementary Appendices B and C, available at http://onlinelibrary.wiley.com/doi/10.1002/acr.23904/abstract). In 16 of 35 studies (46%), whether the translators worked independently was unclear, and in 2 studies they did not work independently, leading to fair and poor ratings, respectively, on the COSMIN checklist. Five studies (14%) reported only a forward translation and an additional 5 studies (14%) reported only 1 forward and 1 backward translation, leading to poor and fair COSMIN

Table 2. Cross-cultural validity using differential item functioning (DIF)*

Study author, year (ref.)	Instrument	Language	Study populations, no.	Findings
Küçükdeveci, 2004 (27)	HAQ DI	Turkish vs. English (UK)	75 Turkish RA patients; comparator: 174 UK RA patients	DIF was seen on 1 of 8 subscales (activity subscale composed of 3 questions, and the DIF on this subscale was not further attributed to any individual tasks). It was unclear whether a formal analysis of the impact of potential bias related to the DIF found was conducted, although authors conclude it is likely negligible.
Oude Voshaar, 2013 (28)	HAQ-II	Dutch vs. English (US)	2 Dutch RA cohorts combined for analysis: A) 472, B) 550; comparator: 18,747 US RA patients	2 items (20%) showed cross-cultural DIF and were more difficult for US patients (standing up from a straight chair and climbing ≥2 flights of stairs). Analysis of impact of potential bias from individual items on total scores deemed negligible.
Oude Voshaar, 2014 (29)	PROMIS physical function item bank	Dutch vs. English (US)	690 Dutch RA patients; 2 comparator groups with 2 independent samples: A) wave 1 US general population (942,995), B) US RA patients (273,280)	A) 25 items (20%) showed cross-cultural DIF (impact on physical function negligible). On 11 items Dutch patients endorsed lower scores (increased difficulty) with use of hands or arms; 12 items were more difficult for US patients, including 5 involving stair climbing. B) 7 of 24 items showed DIF. Analysis of impact of potential bias from individual items on total scores deemed negligible.

* The Health Assessment Questionnaire II (HAQ-II) is a 10-item questionnaire. ref. = reference; HAQ DI = HAQ disability index; RA = rheumatoid arthritis; PROMIS = Patient-Reported Outcomes Measurement Information System.

	Abourazzak, 2008	Bae, 1998	Bruhlmann, 1994	Cardiel, 1993	Chopra, 2012	Ekdahl, 1988	El Meidany, 2003	Esteve-Vives, 1993	Ferraz, 1990	Guillemin, 1992	Hussein, 2008	Islam, 2013	Kirwan, 1986	Koh, 1998	Küçükdeveci, 2004	Matsuda, 2003	Nazary-Moghadam, 2017	Osiri, 2001	Ranza, 1993	Rastmanesh, 2010	Senerdem, 1999	Shakibi, 2012	Shehab, 1998	Studies modified (%)
Dressing																								
Dress yourself, including tying shoelaces and doing buttons?					✔																			4
Shampoo your hair?					✔										✔		✔							13
Arising																								
Stand up from a straight chair?									✔	✔	✔		✔	✔										22
Get in and out of bed?		✔														✔		✔		✔				17
Eating																								
Cut your meat?	✔	✔			✔		✔	✔			✔	✔		✔		✔	✔	✔		✔		✔		57
Lift a full cup or glass to your mouth?	✔							✔																9
Open a new milk carton?	✔			✔	✔			✔	✔	✔	✔	✔	✔	✔			✔	✔						52
Walking																								
Walk outdoors on flat ground?																								0
Climb up five steps?			✔			✔									✔									13
Hygiene																								
Wash and dry your body?								✔				✔		✔										13
Take a tub bath?				✔	✔			✔	✔				✔	✔			✔	✔			✔	✔		43
Get on and off the toilet?					✔															✔		✔	✔	17
Reach																								
Reach and get down a 5 pound object (such as a bag of sugar) from above your head?	✔	✔	✔		✔	✔		✔			✔		✔		✔	✔	✔	✔		✔	✔	✔		65
Bend down to pick up clothing from the floor?																								0
Grip																								
Open car doors?					✔				✔			✔												13
Open previously opened jars?					✔																			4
Turn faucets on and off?														✔	✔				✔					13
Activities																								
Run errands and shop?												✔	✔		✔				✔					17
Get in and out of a car?					✔				✔			✔												13
Do chores such as vacuuming or yard work?		✔			✔	✔	✔	✔			✔	✔	✔	✔			✔	✔	✔		✔	✔		61
Number of the 20 HAQ items modified	4	4	2	1	11	4	2	6	7	2	4	10	7	5	5	3	3	8	2	5	2	5	1	

Figure 2. Items of the original Health Assessment Questionnaire (HAQ) disability index that required modification as part of cross-cultural adaptation. Items requiring modification are indicated by a check mark.

ratings, respectively; 19 studies (54%) did not report the review of the translation by a committee (see Supplementary Appendices B and C, available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.23904/abstract).

During the process of linguistic translation and cultural adaptation, a number of the studies reported modification of the original instrument questions. For example, Figure 2 shows the HAQ DI items that required modification for cross-cultural adaptation in 23 of the 29 HAQ DI studies (6 studies were excluded from this table, because they did not include enough details about translated and modified items). Modifications were required and justified as part of the cross-cultural adaptation process to ensure that each question was not only translated but also culturally relevant to the target population. Of note, 10 of the 23 articles included in Figure 2 modified ≥20% of the items in the original HAQ DI. Two studies modified 50% or more of the items included in the HAQ DI. HAQ DI functional tasks such as "cut your meat," "open a new milk carton," "take a tub bath," and "doing chores" were modified in >40% of these studies. The most commonly modified item (65% of the studies) was converting imperial to metric measurement in "reach and get down a 5-pound object." Car-related activities were changed to bus or even, in 1 study, rickshaw. Toilets often

required specification as to what type of toilet: western, squat, or field toilets. Bathtubs were not used in all cultures and this question was omitted from some translated versions or switched to a shower. Additionally, some cultural modification was required to "cut your meat" if meat was not consumed. In some countries "milk cartons" are not used and the question was changed to "plastic bag of milk." "Do chores such as vacuuming or yardwork" was also modified in some translated versions to other household tasks if vacuuming and yardwork were not frequent tasks.

There were very few studies that examined the cross-cultural validity of the translated instruments in RA populations using IRT with DIF as a formal test of invariance across cultures. These studies are shown in Table 2. Two studies noted DIF between Dutch and US populations for the HAQ-II and PROMIS measures, and a third study found DIF between Turkish and UK populations on the HAQ DI. Küçükdeveci et al (27; 2004) showed invariance across cultures of the Turkish HAQ with absent DIF except for "activities," which were more difficult for the Turkish compared to the UK cohort. Oude Voshaar et al (28; 2013) showed negligible DIF supporting cross-cultural equivalence of the Dutch HAQ-II. Standing up from a straight chair and climbing ≥2 flights of stairs was scored higher for US compared to Dutch cohorts.

Oude Voshaar et al (29; 2014) showed high agreement between total estimates of the Dutch PROMIS compared to the US. Dutch patients had more difficulty with activities that involved the upper extremities, and patients in the US had more difficulty with those activities that involved climbing stairs.

DISCUSSION

This systematic review examines the translation, cultural adaptation, and cross-cultural validity of FSAMs for RA, including the HAQ DI, HAQ-derived measures, and the PROMIS FSAMs. While the initial primary aim of this study was to evaluate the cross-cultural validity of FSAMs, there were unfortunately very few articles that formally assessed this validity in RA (3 of the 39 studies [27–29]). While translation and cultural adaptation alone allow for rapid use of an instrument and serve the local need for monitoring of disease severity and response to treatment, without assessing cross-cultural validity, these measures cannot accurately compare physical function across diverse populations for the purposes of research and clinic trials. Therefore, a major finding of our study was the paucity of studies evaluating the cross-cultural validity of these instruments in RA populations.

In the assessment of cross-cultural validity, when the translated and adapted item shares the same item-response function as the original item, then the item goes beyond linguistic equivalence and cultural relevance and has been shown to work the same way, measuring the same level of function in both cultures. Similarly, the presence of DIF is imperative to know, when comparing scores for translated measures between different groups of participants. For example, knowing that Dutch RA patients (29) had more difficulty with activities that involved the upper extremities and patients in the US with those that involved climbing stairs is helpful when comparing scores from the 2 populations. Differences in these scores may not reflect a difference in function in the populations but instead reflect the underlying difference in functioning of the translated measure and the originator. Here, different Dutch versus US respondents with the same level of physical function responded differently, so scores are not directly comparable, and adjustments might need to be made if studies plan to pool FSAM scores from across different patient populations. Importantly, although DIF was shown in the identified studies, all authors concluded that the effects on overall scores were minimal, supporting the cross-cultural validity of these measures.

The remainder of the studies reported only on the translation and/or cultural adaptation of the FSAM. Overall, we have found that the methodologic quality of the studies describing the translation and/or cultural adaptation of the translated FSAMs is mostly of poor or fair quality according to COSMIN criteria, although many of the studies were conducted before the COSMIN criteria were published. There appeared to be a trend to improvement in study quality, with fewer studies rating poorly on COMSIN criteria after 2009, likely reflecting increasing awareness over time of guidelines highlighting the importance of methodologically sound and transparent methods of translation and cultural adaptation of patient questionnaires (12–18).

The most common shortcoming leading to low COSMIN ratings was lack of any form of pretesting of the translated FSAMs, with only 57% of studies describing any form of pretesting (16 of 37 eligible studies). There were 2 commonly employed methods for pretesting identified in this review: indepth and in-person cognitive interviews, or supplementary questionnaires where participants were asked to rate the comprehensibility of each question and suggest alternative translations. While the first approach is considered superior (26), even the more simplistic approach of supplementary questionnaires allowed for modifications based on participant feedback that improved the ease of interpretability of the instrument prior to its use. Pretesting is an important step in measuring translation, and lack of testing could impact the interpretability of the final translated measure.

While our study represents, to the best of our knowledge, the most comprehensive review on this topic to date, there are some limitations that should be highlighted. First, there may be more translated versions of the included FSAMs than are captured in this current study. The HAQ DI is said to be translated into more than 60 languages and dialects (11). Despite our rigorous search strategy, we found only 22 different translations (not including dialects). There are a few reasons why further translations may have been missed; for example, if translation was not the aim of the study but a part of the methods of an article, and was not specified in the title or abstract, that study could have been missed by our search strategy. Alternatively, an article could have been missed if the translation was done but not published. Last, if the translated measure was published in a non-English language publication or if the study was in a non-RA population, it was not captured in our study.

There have been extensive and cohesive efforts to translate and culturally adapt the PROMIS measures into other languages using rigorous methodology (30), although many of these studies were not included in our results, because they were not conducted in RA populations. The Functional Assessment of Chronic Illness Therapy (FACIT) translation methodology (31) has been adopted for the PROMIS and related measures. FACIT methodology involves the standard steps of iterative forward and back translations, multiple reviews, and cognitive debriefing in the target population. For any new translations, the organization conducts a quality review to ensure harmonization across languages. This process ensures consistency in the translation of similar items within the item bank as well as consistency with previous translations. To date, many translated versions of the PROMIS FSAMs are available for use through HealthMeasures, a US-based national person-centered assessment resource that includes the PROMIS measures (32).

Given the standardized, high-quality approach to translation of the PROMIS measures, the translation and cultural adaptation of translated versions may not always be published. Furthermore, where the information about the process of PROMIS measure translation is published, it may not be in an RA subpopulation, for example, a recently published psychometric evaluation of the German translation of the PROMIS physical function item bank (33). Furthermore, the assessment of the cross-cultural validity of the measures may also have been conducted in populations other than RA and would not have been captured by our literature search. For example, the cross-cultural validity of the PROMIS physical function bank has been examined in >1,000 Dutch patients with chronic pain, and 4 items were flagged for DIF, which negligibly impacted overall scores (34). In contrast, a Spanish translation of the PROMIS physical function item bank was tested in the US general population, revealing that 50 of the 114 items were flagged for important DIF, indicating that English- and Spanish-speaking individuals with the same underlying level of physical function responded differently to these questions (35). In the future, research on the cross-cultural validity of PROMIS physical function measures should be conducted to determine whether disease-specific studies are needed or whether items function similarly in individuals with different chronic diseases within a culture. If there is evidence that items function differently in RA populations, then additional studies should be conducted, given the lack of studies that we identified.

Since the completion of our study, the COSMIN group has published a new risk-of-bias checklist to evaluate patient-reported outcome measures (36). In this new checklist, cultural validity has a broader definition and includes not only ethnically or linguistically different groups but other concepts of culture, including age and sex. The new checklist still defines the evaluation of cross-cultural validity by assessing whether a scale is invariant or whether DIF occurs. In the new COSMIN checklist for assessing cross-cultural validity, the standards for assessing the translation process and for assessing a cross-cultural study were removed, because the authors felt these standards were not a measurement property. We have decided to report on these items, because the results were still informative as highlighted in our study. The COSMIN group also moved the pretest criteria to a different part of the COSMIN checklist under "content validity," with the rationale that "a poor translation process does not necessarily mean that the instrument has a poor cross-cultural validity." While we concur with this statement, we found that reviewing all elements of translation, pretesting, and cross-cultural validity has clearly illustrated important gaps in the methods of translation and testing of widely used FSAMs in RA.

In conclusion, FSAMs have been widely used both in their validated English form and in many translated forms. Although current evidence is limited, that evidence supports the cross-cultural validity of the measures in some cultural settings. Further

investigation should be considered when using these measures for multinational clinical trials and for day-to-day use in practice in settings where cross-cultural validity has not been formally established.

AUTHOR CONTRIBUTIONS

REFERENCES

1. Michaud K, Vera-Llonch M, Oster G. Mortality risk by functional status and health-related quality of life in patients with rheumatoid arthritis. J Rheumatol 2012;39:54–9.

2. Pincus T, Brooks RH, Callahan LF. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. Ann Intern Med 1994;120:26–34.

3. Sokka T, Pincus T. Poor physical function, pain and limited exercise: risk factors for premature mortality in the range of smoking or hypertension, identified on a simple patient self-report questionnaire for usual care. BMJ Open 2011;1:e000070.

4. Cohen JD, Dougados M, Goupille P, Cantagrel A, Meyer O, Sibilia J, et al. Health assessment questionnaire score is the best predictor of 5-year quality of life in early rheumatoid arthritis. J Rheumatol 2006;33:1936–41.

5. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.

6. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.

7. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42:2220–30.

8. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum 2004;50:3296–305.

9. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45 Suppl 1:S3–11.

10. Singh JA, Saag KG, Bridges SL Jr, Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology guideline for the treatment of rheumatoid arthritis. Arthritis Rheumatol 2016;68:1–26.

11. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.

12. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417–32.

13. Bullinger M, Alonso J, Apolone G, Leplege A, Sullivan M, Wood-Dauphinee S, et al. Translating health status questionnaires

and evaluating their quality: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol 1998;51: 913–23.

14. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25:3186–91.

15. Koller M, Aaronson NK, Blazeby J, Bottomley A, Dewolf L, Fayers P, et al. Translation procedures for standardised quality of life questionnaires: the European Organisation for Research and Treatment of Cancer (EORTC) approach. Eur J Cancer 2007;43:1810–20.

16. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract 2011;17:268–74.

17. Koller M, Kantzer V, Mear I, Zarzar K, Martin M, Greimel E, et al. The process of reconciliation: evaluation of guidelines for translating quality-of-life questionnaires. Expert Rev Pharmacoecon Outcomes Res 2012;12:189–97.

18. World Health Organization. Process of translation and adaptation of instruments. 2018. URL: http://www.who.int/substance_abuse/research_tools/translation/en/.

19. 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;19:539–49.

20. Consensus-based Standards for the selection of health Measurement Instruments Initiative (COSMIN). COSMIN. 2018. URL: www.cosmin.nl.

21. Regnault A, Herdman M. Using quantitative methods within the universalist model framework to explore the cross-cultural equivalence of patient-reported outcome instruments. Qual Life Res 2015;24:115–24.

22. Siemons L, Krishnan E. A short tutorial on item response theory in rheumatology. Clin Exp Rheumatol 2014;32:581–6.

23. PRISMA. PRISMA: transparent reporting of systematic reviews and meta-analyses. 2015. URL: http://www.prisma-statement.org/.

24. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115–23.

25. 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.

26. Collins D. Pretesting survey instruments: an overview of cognitive methods. Qual Life Res 2003;12:229–38.

27. Küçükdeveci AA, Sahin H, Ataman S, Griffiths B, Tennant A. Issues in cross-cultural validity: example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. Arthritis Rheum 2004;51:14–9.

28. Oude Voshaar MA, Glas CA, ten Klooster PM, Taal E, Wolfe F, van de Laar MA. Crosscultural measurement equivalence of the health assessment questionnaire II. Arthritis Care Res (Hoboken) 2013;65:1000–4.

29. Oude Voshaar MA, ten Klooster PM, Glas CA, Vonkeman HE, Taal E, Krishnan E, et al. Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. PLoS One 2014;9:e92367.

30. Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, et al. The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. Health Qual Life Outcomes 2013;11:210.

31. Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof 2005;28:212–32.

32. Health Measures. PROMIS. 2018. URL: http://www.healthmeasures.net/explore-measurement-systems/promis.

33. Liegl G, Rose M, Correia H, Fischer HF, Kanlidere S, Mierke A, et al. An initial psychometric evaluation of the German PROMIS v1.2 physical function item bank in patients with a wide range of health conditions. Clin Rehabil 2018;32:84–93.

34. Crins MH, Terwee CB, Klausch T, Smits N, de Vet HC, Westhovens R, et al. The Dutch-Flemish PROMIS physical function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol 2017;87:47–58.

35. Paz SH, Spritzer KL, Morales LS, Hays RD. Evaluation of the Patient-Reported Outcomes Information System (PROMIS(®)) Spanish-language physical functioning items. Qual Life Res 2013;22:1819–30.

36. Mokkink LB, de Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27: 1171–9.