Received: 25 May 2020 | Revised: 22 August 2020 | Accepted: 18 September 2020

DOI: 10.1111/sms.13854

WILEY

ORIGINAL ARTICLE

How to translate and locally adapt a PROM. Assessment of cross-cultural differential item functioning

Michael R. Krogsgaard[1] | **John Brodersen**[2,3] | **Karl B. Christensen**[4] | **Volkert Siersma**[2] | **Jonas Jensen**[1] | **Christian Fugl Hansen**[1] | **Lars Engebretsen**[5] | **Håvard Visnes**[6,7] | **Magnus Forssblad**[8] | **Jonathan D. Comins**[1,2]

[1]Section for Sports Traumatology M51, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark

[2]The Research Unit for General Practice and Section of General Practice, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[3]Primary Health Care Research Unit, Region Zealand, Sorø, Denmark

[4]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[5]Orthopedic Clinic, Oslo Sports Trauma Research Center, University of Oslo Medical School, Oslo, Norway

[6]Norwegian National Knee Ligament Registry, Department of Orthopedic Surgery, Haukeland University Hospital, Bergen, Norway

[7]Oslo Sports Trauma Research Center, Norwegian School of Sports Sciences, Oslo, Norway

[8]Department of Molecular Medicine and Surgery, Stockholm Sports Trauma Research Center, Karolinska Institute, Stockholm, Sweden

Correspondence
Michael R. Krogsgaard, Section for Sports Traumatology M51, Bispebjerg and Frederiksberg Hospital, Bispebjerg Bakke 23, DK-2400 Copenhagen NV, Denmark.
Email: michael.rindom.krogsgaard@regionh.dk

Translating patient-reported outcome measures (PROMs) can alter the meaning of items and undermine the PROM's psychometric properties (quantified as cross-cultural differential item functioning [DIF]). The aim of this paper was to present the theoretical background for PROM translation, adaptation, and cross-cultural validation, and assess how PROMs used in sports medicine research have been translated and adapted. We also assessed DIF for the Knee Injury and Osteoarthritis Outcome Score (KOOS) across Danish, Norwegian, and Swedish versions. We conducted a search in PubMed and Scopus to identify the method of translation, adaptation, and validation of PROMs relevant to musculoskeletal research. Additionally, 150 preoperative KOOS questionnaires were obtained from the Scandinavian knee ligament reconstruction registries, and cross-cultural DIF was evaluated using confirmatory factor analysis and Rasch analysis. There were 392 studies identified, describing the translation of 61 PROMs. Ninety-four percent were performed with forward-backward technique. Forty-nine percent used cognitive interviews to ensure appropriate wording, understandability, and adaptation to the target culture. Only two percent were validated according to modern test theory. No study assessed cross-cultural DIF. One KOOS subscale showed no cross-cultural DIF, two had DIF with respect to some (but not all) items, and thus conversion tables could be constructed, and two KOOS subscales could not be pooled. Most PROM translations are of undocumented quality, despite the common conclusion that they are valid and reliable. Scores from three of five KOOS subscales can be pooled across the Danish, Norwegian, and Swedish versions, but two of these must be adjusted for DIF.

KEYWORDS

cognitive interview, construct validity, cultural adaption, data pooling, differential item functioning, knee ligament reconstruction registry, PROMs, translation

1 | INTRODUCTION

A common reason for translating and adapting patient-related outcome measures (PROMs) from one language to another is that a specific PROM is needed for a study but does not exist in the local language. If a PROM has been developed with help from relevant patient groups, using valid methods, so it has content relevance and coverage for the patients in the planned study, then this is a good reason to translate and adapt the existing PROM instead of developing a new one. This is easier and less time-consuming.

In other cases, there is a desire to conduct studies across countries, languages, or cultures, for instance in multi-center trials involving different countries or trials in countries where there is more than one national language. Also, international clinical databases need the same outcome measures in all the participating countries, so data can be pooled or compared, and this includes relevant PROMs. There is an increasing need in relation to planning and financing in health policy to be able to compare clinical outcomes from different countries or cultural groups. PROMs are important in this context, which emphasizes that measurement must be independent of language and culture.

To adapt a PROM to a new language or culture is not trivial. Even for languages that are spoken by many people globally across different countries, such as Spanish, English, and Arabic, the same basic language can have quite varied versions, as the habits and cultures of the different countries can diverge substantially. The same word or expression can carry different connotation and meaning across the different countries, or objects can be described by different words in the same language, dependent on culture or geography. For example, "braces" in the United Kingdom (UK) are called "suspenders" in the United States (US), where "braces" are used to straighten teeth.

Also, life conditions can be very different within language areas, dependent on socioeconomic, religious, and cultural conditions and are often very different between countries. Therefore, the content of the items in a PROM may not have the same meaning or importance when it is translated to a new culture.

All these issues create methodological challenges when a PROM is translated and adapted to a new language and culture.

There are several ways to conduct translation and adaption, and there is evidence that a rigorous and multistep procedure leads to a better translation and adaption.[2]

Once a PROM has been translated and adapted, it should be confirmed that it measures in the same way (invariantly) for all persons. Even within the same language and culture, items can function differently dependent on for instance gender or age, and this is called differential item functioning (DIF).[3,4] This is probably even more pronounced between countries and cultures (cross-cultural DIF), for instance do Norwegians understand and respond to items in the same way as Americans? If results are compared between cultures or countries, or if data from several countries are pooled, items that have cross-cultural DIF introduce a systematic bias that will give respondents in different countries a different score, even though their condition is the same. For example, it was demonstrated by comparing results from the three Scandinavian knee ligament reconstruction registries that Danish patients have significantly lower scores in the

Case

Three strategies (debridement, microfracture, and no treatment) to handle full-thickness lesions of knee hyaline cartilage were evaluated by identifying patients with a knee ligament reconstruction and a cartilage lesion in the Norwegian and Swedish National Knee Ligament Registries. The outcome two years after surgery was the Knee Injury and Osteoarthritis Outcome Score (KOOS). Linear regression analyses were used to evaluate the effect of debridement and microfracture on the domain scores of KOOS[1].

No significant effects of debridement were found on any of the KOOS subscales at two-year follow-up compared to no treatment. Microfracture treatment was associated with significantly worse scores compared to no treatment at two-year follow-up in the KOOS Sport and Recreation and Knee-related Quality of Life subscales. For the remaining KOOS subscales of Pain, Symptoms, and Activities of Daily Living, there were no significant effects of microfracture.

It was concluded that microfracture of concomitant full-thickness cartilage lesions showed adverse effects on patient-reported outcomes at two-year follow-up after ACL reconstruction. Debridement of concomitant full-thickness cartilage lesions showed neither positive nor negative effects on patient-reported outcomes at two-year follow-up after ACL reconstruction.[1]

Comment: The psychometric properties of the Norwegian and Swedish versions of KOOS have not been compared in a joint data set with individuals from both countries, so it is not known, if data from the two cohorts can be directly pooled. Whether KOOS functions differently across countries can be tested in a pooled dataset. If items or scales function differently between countries, this can often be adjusted for by using conversion tables derived from pooled data sets.

KOOS domain "Symptoms" compared to their Norwegian and Swedish counterparts, both preoperatively and postoperatively.[5] Therefore, cross-cultural DIF can be suspected for items in this domain.

The presence of cross-cultural DIF is of course most important if data from different countries or cultures are pooled into one dataset. This is typically done in international databases or when national clinical databases are pooled, but also randomized multicentre studies and studies including cohorts in different countries can be affected by cross-cultural DIF, like the Delaware-Oslo cohort of ACL patients.[6,7]

1.1 | The theoretical background

In most cases, PROMs are developed in one language and culture and then translated and adapted to other languages and settings. The most commonly used PROMs in sports science were all developed within the Western culture.[8] The main and most important objective of the translation and adaptation process of a PROM across settings is to transfer the meaning of each item and construct encompassed in the PROM from the original language and culture into another language and culture. This involves transfer of the wording as well as the relevance of each item.

There are four criteria, which must be considered for the translated PROM, as defined by Beaton[9]:

1. Semantic equivalence, meaning grammatical, and vocabulary equivalence with the original PROM. Ambiguous wordings are avoided (ie, the translated words must have one meaning and be understandable to everyone).
2. Idiomatic equivalence. Some expressions are idioms, meaning that the words themselves give no understanding of the expression. An example is "feeling downhearted and blue" (from Short Form 36 (SF-36)). Idioms must be reworked beyond translation, but for some idioms, there is no equivalent expression in target languages.
3. Experiential equivalence, meaning that some activities are not the same in the local setting and must be replaced by something equivalent. An example is that skiing was replaced by surfing in the translation of a PROM from American English to Brazilian Portuguese.[10]
4. Conceptual equivalence, meaning that specific concepts (for instance "family," "work," and "leisure time") may have very different meanings in different cultures, which can result in different answers.

It is generally recommended that questionnaires can be understood by the equivalent of a 12 year old (Grade 6 reading level),[9] but the importance of this is of course dependent on the target population and its educational level. This can be a problem in countries, where a larger proportion of inhabitants do not have an educational level past Grade 6.

1.1.1 | Translation and cultural adaption

The first part of the process to translate a PROM into a local language is of course to translate the wording of the items and the instruction. The two most accepted methods are somewhat different: forward-backward translation and dual-panel translation. The steps are described in Boxes S1 and S2.

Of the two methods, the most frequently used is *forward-backward translation*, described in detail by Beaton.[9] With this method, the translation is sometimes performed by linguistic experts (eg, professional translators) or healthcare professionals, and thus, there is a risk that the wording will not be in common lay language and thereby has suboptimal meaning or readability for the majority of the general population. This can only be addressed by conducting some kind of cognitive interviewing or field test of the understandability of the wording after the forward-backward translation has been conducted to ensure that meaning is not lost and that the translated version of the PROM is understandable for lay people.[9] As PROMs in most cases are completed by laypersons who are patients, cognitive interviewing regarding the wording should primarily be performed with laypersons. Healthcare professionals tend to use professional phrases, and patients tend to focus more on their disease(s) and thereby the subject matter in the PROM than on the actual language, meaning, and understandability, and neither of these groups are optimal for cognitive testing of the wording (the language).

However, patients with the condition that the PROM is meant to cover can participate in cognitive testing of the understandability of the translated PROM—does the wording make sense for the subjective understanding of the condition? This can be necessary, as a translation by professional translators can be linguistically correct, but not meaningful for the target group. This means that after the forward-backward translation has been carried out, the PROM needs to be field-tested through cognitive interviews for understandability, and, if necessary, modified.

Conversely, the main purpose of the *dual-panel translation and adaptation* method is to ensure the quality of the translation during the translation process itself[11] (Box S2). The primary translation is made in a group of bilingual persons, and the wording is discussed (and possibly modified) until the group agrees that meaning of the wording in the original version is covered in the translated version. The second panel includes a lay panel of 3-5 local persons, who in plenum can discuss the wording and modify the items that have been proposed by the first bilingual panel. So, if the dual-panel method is used, it is not necessary additionally to

test the translated version for wording or understandability, as this is already part of the method.

Preferably, the researcher involved in developing the original PROM can be part of the entire translation and adaptation process and help ensure that the meaning of the items and constructs are kept in the translation process across the settings.[11]

1.1.2 | Assessing the psychometric properties of the translated PROM

Regardless of which translation and adaptation method is used, an equally important aspect is to conduct psychometric analyses to confirm the construct validity of the PROM scales in the new setting and ideally whether there is DIF across the settings (ie, across the two versions).[4] Does the PROM measure the same single construct, or multiple constructs, in both settings, and do people in both settings interpret the items in the same way? Language DIF is in particular important to consider when comparing data and results from different countries, for instance in relation to publications of combined data from several countries (eg, from National clinical databases such as knee ligament reconstruction registries, and arthroplasty registries). However, when psychometric properties are tested, it is usually only performed on data collected from one country, and thus, cross-cultural analyses of the psychometric properties between the original and the translated measure are not addressed.[4] This is suboptimal if results are compared between countries. When PROM data are analyzed in pooled data sets with data from more than one country, simple adjusting for the effect of country in a regression model is not sufficient. Consider the following analogy: A multi-center study measures the primary outcome as changes in temperature. Some centers use Celsius while others use Fahrenheit. Adding an effect of country in your regression model will not yield a correct analysis. However, knowing how to translate from one temperature scale to the other will enable you to do a valid analysis. Therefore, conversion tables are required.

The optimal procedure of cross-cultural analysis is to evaluate validity in each language version separately and subsequently pool collected data and assess measurement invariance and DIF relative to language for each domain score in the pooled data set. In this way, it is possible to reveal whether persons with the same overall score on the remaining items systematically give different responses to the item being tested. If the difference in mean item scores for an item with DIF for the pooled scores (ie, the combined data) is uniform along the scale (as measured by the total score), then this difference can be adjusted across the settings, so long as fit to a measurement model is maintained.[3] If this is the case, the item displays DIF across country, language, and culture. Once DIF has been identified, it can be compensated for using conversion tables, when data are reported. Measurement invariance can be tested using multiple groups confirmatory factor analysis (CFA),[12] while DIF is most easily tested using item response theory (IRT). DIF can best be explained using the item location. For example, in a scale that measures the impact of knee function on quality of life, an item that assesses whether the respondent is able to go cross-country skiing would have a different location (ie, level of difficulty on the scale) for Swedes and Norwegians (who have a long tradition for skiing regularly) compared to Danes (who mainly go skiing during vacations). It would be expected that a small proportion of Danish respondents, but a larger proportion of Swedes and Norwegians, would report this to have an impact on health-related quality of life. Since the ordering of all items in terms of level of difficulty included in a scale can be determined using IRT models, this provides a way to test items in scales for DIF in relation to country, language, and culture.[3] Such analyses for unidimensionality and DIF can provide robust evidence that the same constructs are actually measured in the same way across different borders and that this is done invariantly.[3] Results of PROM scores that are pooled from several countries can be different, dependent on whether DIF has been compensated for or not.

1.2 | Hypotheses and aims

It is stated in most articles reporting translation and adaption of a PROM that it was found to be a valid and reliable measurement tool in the translated version. However, it is not known to which extent translation, adaptation, and validation of versions in languages other than the original PROMs in sports in fact has been performed optimally. It was hypothesized that for a majority of PROMs used in sports research optimal methods had not been employed in the adaptation and validation of translated versions. Furthermore, it was hypothesized that calculation of local DIF and cross-cultural DIF was generally not performed.

In relation to the Scandinavian knee ligament reconstruction registries, it can be relevant to pool data from the three countries (Norway, Sweden, and Denmark). However, it has never been assessed whether there is cross-cultural DIF for the main outcome, KOOS. It was hypothesized that there may be cross-cultural DIF between the local Scandinavian versions of KOOS and that this can be compensated for, when pooled data are reported.

The aims were therefore twofold:

1. To study how translation, adaptation, and validation were performed in the local versions of the most commonly used and relevant PROMs in Sports. These comprised 61 PROMs which had been identified from searches

in PubMed 2011-2020, being either commonly used (more than three times during this time period), used in randomized studies on musculoskeletal conditions or being the only PROM for a specific musculoskeletal condition of relevance. Translated versions of these 61 PROMs were searched for in PubMed and Scopus. This is described in detail elsewhere.[8]
2. To assess cross-cultural DIF in the questionnaire KOOS between Denmark, Sweden, and Norway.

2 | METHODS

2.1 | Aim 1

All published translated versions of the 61 PROMs that were identified in[8] were analyzed.

The quality indicators for translation and adaptation of a PROM for use in another country, language, or culture were defined by three components:

1. *Translation and adaptation:* Has the meaning of the items and constructs in the PROM been adequately transferred from the original language and culture to the other language and culture?
2. *Validation of the construct of the translated scale:* Has a test of unidimensionality and DIF of the scale(s), optimally using IRT models, been conducted?
3. *Functioning of the translated PROM compared to the original version:* Has a test of item ordering in scale(s), using IRT models, been conducted, both separately for the countries and with the data from the different countries combined (ie, are the ordering and locations consistent across countries)? Has a cross-cultural DIF analysis been conducted with data from the different countries combined?

Validation of the construct(s) was not included in the analyses for this study, as this has been assessed elsewhere.[8] Also, assessment of development of the original version has been covered in.[8]

Details of the analyses are supplied in the supplementary materials ("Details of recorded information").

2.2 | Aim 2

To assess cross-cultural DIF for KOOS in Denmark, Norway, and Sweden, data from questionnaires completed preoperatively were obtained from National knee ligament reconstruction registries in each country. From each registry responses from 75 women and 75 men, aged 18-37 years, between 2016 and 2018 where included. Validity was evaluated using CFA and Rasch models and the hypothesis of measurement invariance that the latent variables are understood and measured in the same way across countries,[13] and absence of cross-cultural DIF was tested using multiple groups CFA by the latest available guidelines[14] and graphical Rasch models.[15] The R package lavaan[16] and the software package DIGRAM[17] were used.

For all subscales, the following analyses were considered: First, validity in each country was assessed using CFA and Rasch analysis, controlling the type I error rate using the false discovery rate.[18] Second, the fit of a multiple groups CFA models with configural invariance and of graphical Rasch models were evaluated.

For subscales where these basic validity requirements were met, multiple group CFA models and graphical Rasch models with invariance were fitted. Subscales where these restricted models fitted were categorized as having measurement invariance and no DIF. For subscales where this was not the case, models with partial invariance were applied to identify items with DIF. Model fit is evaluated using chi-square test for CFA models and Andersens conditional likelihood ratio test for Rasch models.[19]

For subscales where models with partial invariance could be fitted to the data, conversion tables are reported.

3 | RESULTS

3.1 | Aim 1

3.1.1 | Translation

Of the analyzed 392 PROM studies, direct translation by the researcher, with no formal procedure to secure quality, had been performed in 16. In 368 PROM studies (94%), the forward-backward method was used, and one study used the dual-panel method (Tables S1-S9). In 6 cases, the method of translation had not been described.

3.1.2 | Language adaption

Among the 391 PROMs that had not been translated by the dual-panel method, wording had been discussed through individual interviews in 192 (49%) (Tables S1-S9). In 120 cases (31%), the understandability was tested by analyses of filled out questionnaires but without interviews. In 61, the wording had not been discussed and in 16 it was not described if wording had been discussed.

3.1.3 | Content adaption

In 291 (74%) of the translated PROMs, patients had been involved in testing relevance and understandability, while

this was not the case in 80 and not described in 19 cases (Tables S1-S9). In 194 cases (49%), the pre-version of the PROM had been modified after testing, while no changes had been applied in 168 cases.

3.1.4 | Unidimensionality

In 11 cases (3%), unidimensionality had been assessed for the translated version, in no cases for the original and the translated versions individually, and in no cases for the pooled data set (Tables S1-S9).

3.1.5 | Cross-cultural DIF

DIF had not been assessed for the local PROM in any case. Cross-cultural DIF had been assessed in one case (for The Western Ontario and McMaster Universities Osteoarthritis Index [WOMAC]) but not in relation to translation (Tables S1-S9).

3.2 | Aim 2

Fit indices for models where no items were restricted to be equal across countries (sometimes called "configural invariance" models) showed poor fit for all subscales except Quality of Life (QoL) (results not shown). Adjustment for multiple testing (five subscales in three countries using two different methods yielding 30 statistical tests) was used. Additional analyses using models with correlated error terms/local response dependence showed adequate fit for all subscales except Activities of Daily Living (ADL). No model with correlated error terms/local response dependence fitted this subscale.

Since there is no point in evaluating cross-cultural validity when there is no evidence of validity in any of the three countries, the question of cross-cultural validity was addressed for the four other subscales only. Fit indices for multiple group analyses for these are reported in Table 2. For the ADL subscale that did not meet validity requirements in any of the countries, evaluation of cross-cultural validity was meaningless.

Fit indices for models where no items were restricted to be equal across countries (sometimes called "configural invariance" models) showed adequate fit for the QoL subscale only (results not shown). Including local dependence (correlated error terms) yielded models with adequate fit (results not shown).

Fit indices for models where all items were restricted to be equal across countries (sometimes called "scalar invariance" models) showed adequate fit for the QoL subscale only (results not shown). For the three subscales Pain, Symptoms, and Sport, we used multiple groups CFA and graphical Rasch models in an attempt to identify models where some, but not all items were restricted to be equal across countries (sometimes called "partial invariance" models). The items, which are not restricted, are the items that have cross-country DIF. For the Pain subscale, the items P2 and P7 showed DIF, for the Symptoms subscale all items showed DIF, and for the Sport subscale the item Sp4 showed DIF (Table 1). This means that for the Pain subscale and the Sport subscale conversion tables can be constructed (Table 2).

In summary, the assessment of cross-cultural DIF across Denmark, Norway, and Sweden for the KOOS subscales yielded different results for the five subscales. The ADL subscale did not show construct validity in any of the three countries, making evaluation of cross-cultural validity meaningless. The Symptoms subscale was valid in all countries, but all items displayed evidence of DIF. As no items are on the same metric for this domain, translation from the metric of one country to the metric of another country is not possible. The Pain and Sport subscales were valid in all countries, but they had DIF with respect to some (but not all) items. As the items in these two domains without DIF are on the same metric, translation from the metric of one country to the metric of another country can be based on these, and conversion tables could be constructed. The QoL subscale was valid in all countries with no evidence of DIF, and therefore, scores

TABLE 1 Evaluation of models with partial invariance

		CFA validation			Rasch validation		
KOOS subscale	DIF items	Chi-square	DF	*P*	Chi-square	DF	*P*
Pain	P2, P7	109.5	89	.070	129.5	106	.0602
Symptoms	All						
Sport	Sp4	31.8	31	.425	91.3	71	.0529
QoL	None	20.0	20	.459	28.0	20	.1098

Note: All models include local dependence/correlated error terms. For the Symptoms subscale, no differential item functioning (DIF) equating was possible because all items showed DIF. CFA, Confirmatory factor analysis; KOOS, the Knee injury and Osteoarthritis Outcome Score.

TABLE 2 Conversion tables for adjusting for cross-cultural differential item functioning (DIF)

KOOS pain subscale			KOOS sport subscale		
Denmark	**Norway**	**Sweden**	**Denmark**	**Norway**	**Sweden**
0.0	0.0	0.0	0	0.0	0.0
3.7	3.8	2.3	5	5.0	5.3
7.4	7.6	5.2	10	9.8	10.4
11.1	11.2	8.8	15	14.5	15.4
14.8	14.8	12.8	20	19.2	20.3
18.5	18.3	16.9	25	24.0	25.1
22.2	21.7	21.1	30	28.8	29.9
25.9	25.2	25.3	35	33.6	34.6
29.6	28.6	29.5	40	38.5	39.2
33.3	32.1	33.8	45	43.4	43.8
37.0	35.7	38.0	50	48.2	48.3
40.7	39.3	42.1	55	53.1	52.8
44.4	42.9	46.1	60	57.8	57.3
48.1	46.6	49.9	65	62.6	62.0
51.9	50.3	53.6	70	67.5	66.9
55.6	54.0	57.2	75	72.5	72.3
59.3	57.7	60.8	80	77.7	77.8
63.0	61.4	64.3	85	82.9	83.4
66.7	65.0	67.7	90	88.1	88.8
70.4	68.6	71.1	95	93.2	94.1
74.1	72.2	74.4	100	100.0	100.0
77.8	75.7	77.7			
81.5	79.2	80.9			
85.2	82.7	84.2			
88.9	86.4	87.6			
92.6	90.4	91.2			
96.3	94.9	95.3			
100.0	100.0	100.0			

from this subscale for the different countries can be pooled with no conversion.

The conversion table (Table 2) can be used to translate KOOS scores of the Pain and Sport subscales from one country to the metric of the corresponding KOOS subscales score in the other two of the three Scandinavian countries. For example, a Danish patient scoring (2,3,3,1,2) on the five items in the Sport subscale has a score of 50 for the subscale (the mean item score is divided by four and the result is transformed linearly to a zero to 100 scale, 100 indicating no problems and 0 indicates extreme problems, according to the instructions for KOOS). If the score from this patient is compared to or pooled with scores from Norwegians or Swedes, the score must be translated to 48.2 and 48.3, respectively. In a pooled dataset from all the three Scandinavian countries, one country is chosen as reference, and scores from the two other countries are transformed according to Table 2 before they are pooled.

4 | DISCUSSION

4.1 | Aim 1

This study showed that almost all of PROMs had been translated by the forward-backward method based on the instructions described by Beaton et al in 2000,[9] to which almost all authors referred. About half of the translations had followed the instructions regarding translation and cultural adaption in detail, which is better than hypothesized. However, for the

vast majority construct validity had not been assessed by the most adequate methods (modern test theory models), which reduces confidence in the measurement properties.

This shows that the conclusion in most of the 392 manuscripts: "The translated PROM is a valid and reliable measurement tool" would not necessarily be correct, if thorough translation, adaptation, and validation had actually been performed by optimal methods. The better methods, the higher risk there is to find that the PROM is not reliable and valid. Therefore, instead of referring to the conclusion in the translation-manuscript when the choice of PROM for a study is argued for, authors should describe the methods that had been used for translation, adaption, and validation and search literature for additional assessments. There are several examples of translations, which have been assessed as reliable and valid using classical test theory methods only, that have been shown not to be valid when tested using modern test theory—and this should of course be accounted for in the study article.

A surprising but potentially serious problem that this study has identified is that for several PROMs that had been developed in patient populations with a mother tongue which was not English, an English version of the questionnaire was published with the development article, but with no documentation that it had been translated through any controlled process or been adapted in an English speaking country. As these English versions have been basis for the majority of other translations of these PROMs, the validity of the translated versions can, in principle, be questioned. This is the case for the Copenhagen Hip and Groin Outcome Score (HAGOS), the Foot and Ankle Outcome Score (FAOS), and The Achilles Tendon Total Rupture Score. The 5 domains in KOOS and the Hip dysfunction and Osteoarthritis Outcome Score (HOOS) consist of 3 domains from the WOMAC, which were developed in a community of Canadian-English speaking patients, and 2 domains that were developed in a Swedish speaking population, but there is no documentation that WOMAC had been thoroughly translated to Swedish or the two other domains had been thoroughly translated into English. KOOS and HOOS were originally validated in a community of Swedish speaking patients. This means that there is no documented validity of the English versions of KOOS and HOOS, and the Swedish version is questionable, as the process of translation to Swedish of 3 of 5 domains has not been documented. KOOS-Child was developed in a Swedish speaking community, and there is no documentation that the English version is based on a thorough translational and cultural adaptation process. The Achilles Tendon Total Rupture Score was also developed in Swedish, but how translation into the English version that was published in the development article had been performed is not documented. Nine of the 12 translations of this PROM have been made from the English version. The Forgotten Joint Score was developed and validated in a German speaking community, but the English version (from which 5 of 7 translations have been made) has not been documented. The Kujala Score (Anterior Knee Pain Scale) was developed in a Finnish setting, but there is no documentation of the translation to English (from which 9 of 10 translations were made). The Lysholm score was developed in Swedish, and it is not documented how it was translated into English (from which 4 of 6 published translations were made).

In addition to the translations that were identified for this study through academic search strings, there is a large number of translated versions, which have either not been documented or have only been published in gray literature. As an example, there are 51 versions of KOOS, 14 versions of HAGOS, 25 versions of HOOS, 17 of FAOS, and 7 versions of KOOS-Child available (as of January 1, 2020) from www.koos.nu, whereas the respective numbers of identified, published translations are 19, 4, 13, 11, and 2. This shows that it is essential that reports on translation and adaption are actually peer reviewed and published.

It is rare that a PROM is developed simultaneously in different languages and settings. This has been described for KOOS, KOOS-Child, and the Functional Assessment Scale for Acute Hamstring Injuries (FASH). The latter was developed in a Greek community and translated into German and French by the forward-backward method.[20] Even though the process is not described in all details, this has resulted in three valid PROMs. However, it is not a simultaneous development as only Greek patients participated in the development of items. KOOS is a mixture of subscales that were developed in Canada (3 domains) and in Sweden (2 domains) but not simultaneously. So, there are no examples related to musculoskeletal conditions of PROMs developed simultaneously in difference countries or cultures. This would be an optimal method to develop PROMs for patients with rare diseases, for instance children with ACL-rupture, as it is difficult to involve enough patients for development in one country.

A very thorough guide to forward-backward translation and cultural adaption is available in Wild D et al.[21]

4.2 | Aim 2

When PROM data combined from several countries are published, it is a general measure of quality to know, if there is cross-cultural DIF, and if there is, that this DIF is corrected for, before data are pooled. This was first suggested in 2004,[22] but it has not been assessed for PROMs that are relevant for musculoskeletal research.

For KOOS, this study showed that data can be pooled from 1 of the 5 subscales without conversion and for 2 subscales if scores are corrected for cross-country DIF by conversion. For 2 subscales, pooling of data is not meaningful. This is relevant when data from National clinical databases

from several countries are published, or when data from studies in different countries are pooled. There are no examples within sports research where cross-country DIF has been considered in studies where results from several language areas are represented. For observational studies comparing different conditions or treatments (like the study in the opening case of this article), the error that cross-country DIF can introduce depends on the distribution of the conditions/treatments between countries. If for instance one treatment is tradition in one country and another treatment in the second country, comparison of the treatment results is affected by cross-country DIF. For randomized, controlled studies, where allocation to treatment arms is made separately in each country, the means of outcome in the two treatment arm are affected equally by a cross-country DIF, but the variation in the pooled data might increase, if cross-country DIF is not compensated for. If, however, allocation is made for the complete cohort, treatments may not be distributed evenly in each country, and a cross-country DIF may affect the mean of the outcomes and thereby the assessment of a possible difference in outcome of the two treatments. This could be the case for an international multicentre study with a central computer for allocation.

5 | CONCLUSION

About half of the PROMs were translated and adapted by accepted methods. However, the vast majority of translated PROMs have not been validated optimally and are therefore of questionable quality, despite the common individual conclusion of the actual PROM being a valid and reliable measurement tool. There is differential item functioning (DIF) between Denmark, Norway, and Sweden in relation to many items of KOOS, meaning that if data are pooled or compared between countries, this should be corrected for. For two subscales of KOOS, pooled data are not meaningful.

6 | PERSPECTIVES

Ideally, all translated and adapted PROMs should be produced according to standard principles, and in cases where this has not been done, it can be considered to re-translate the PROM. It can be considered for PROMs that have not been validated by modern test theory model methods to re-validate, for instance by use of already existing data. The methods for translation, adaption, and validation should always be described in detail, when results obtained by translated PROMs are published, and if optimal methods have not been used, the implications for the results should be discussed. If PROM scores from different countries are compared or pooled, it should be known whether there is cross-country DIF, and this can be assessed during the process of translation and cultural adaption. Data should be converted before pooling, if there is cross-country DIF.

CONFLICTS OF INTEREST

All authors declare that they have no conflicts of interest in relation to this manuscript.

ORCID

Michael R. Krogsgaard https://orcid.org/0000-0002-9976-4865
John Brodersen https://orcid.org/0000-0001-9369-3376
Karl B. Christensen https://orcid.org/0000-0003-4518-5187
Volkert Siersma https://orcid.org/0000-0003-1941-2681
Christian Fugl Hansen https://orcid.org/0000-0003-4769-4014
Jonathan D. Comins https://orcid.org/0000-0002-8284-114X

REFERENCES

1. Røtterud JH, Sivertsen EA, Forssblad M, et al. Effect on patient-reported outcomes of debridement or microfracture of concomitant full-thickness cartilage lesions in anterior cruciate ligament-reconstructed knees: a nationwide cohort study from Norway and Sweden of 357 Patients With 2-Year Follow-up. *Am J Sports Med*. 2016;44:337–344.
2. Acquadro C, Conway K, Hareendran A, Aaronson N. Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value Health*. 2008;11(3):509–521.
3. Brodersen J, Meads DM, Kreiner S, Thorsen H, Doward L, McKenna SP. Methodological aspects of differential item functioning in the Rasch model. *J Med Econ*. 2007;10(3):309–324.
4. Holland PW, Wainer H (Eds). *Differential Item Functioning*. New York: Laurence Erlbaum Associates; 1993.
5. Granan LP, Forssblad M, Lind M, Engebretsen L. The Scandinavian ACL registries 2004–2007: baseline epidemiology. *Acta Orthop*. 2009;80:563–567.
6. Grindem H, Wellsandt E, Failla M, Snyder-Mackler L, Risberg MA. Anterior cruciate ligament injury—who succeeds without reconstructive surgery? The Delaware-Oslo ACL Cohort Study. *Orthop J Sports Med*. 2018;6:2325967118774255.
7. Capin JJ, Failla M, Zarzycki R, et al. Superior 2-year functional outcomes among young female athletes after ACL reconstruction in 10 return-to-sport training sessions: comparison of ACL-SPORTS randomized controlled trial with Delaware-Oslo and MOON Cohorts. *Orthop J Sports Med*. 2019;7:2325967119861311.
8. Hansen CD, Jensen J, Siersma V, Brodersen J, Comins JD, Krogsgaard MR. A catalogue of PROMs in sports science - quality assessment of PROM development and validation. *Scand J Med Sci Sports*. 2020.
9. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*. 2000;25:3186–3191.
10. Metsavacht L, Leporace G, Riberto M, Sposito MMdM, Batista LA. Transation and cross-cultoral adaption of the Brazilian version

of the International Knee Documentation Committee subjective knee form. *Am J Sports Med*. 2010;38:1894–1899.
11. Swaine-Verdier A, Doward LC, Hagell P, Thorsen H, McKenna SP. Adapting quality of life instruments. *Value Health*. 2004;7:S27–S30.
12. Jöreskog KG. Simultaneous factor analysis in several populations. *Psychometrika*. 1971;36:409–426.
13. Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*. 1993;58:525–543.
14. Svetina D, Rutkowski L, Rutkowski D. Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M*plus* and the lavaan/semTools packages. *Struct Equa Model*. 2020;27:111–130.
15. Kreiner S, Christensen KB. Graphical Rasch models. In: Mesbah M, Cole FC, Lee MT, eds. *Statistical methods for quality of life studies*. Boston, MA: Springer; 2002:187–203.
16. Rosseel Y. lavaan: An R Package for structural equation modeling. *J Stat Softw*. 2012;48:1–36.
17. Kreiner S, Nielsen T. *Item analysis in DIGRAM 3.04: Part I: Guided tours*. Copenhagen, Denmark: University of Copenhagen. 2013.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc*. 1995;57:289–300.
19. Andersen EB. A goodness of fit test for the rasch model. *Psychometrika*. 1973;38:123–140.
20. Malliaropoulos N, Korakakis V, Christodoulou D, et al. Development and validation of a questionnaire (FASH–Functional Assessment Scale for Acute Hamstring Injuries): to measure the severity and impact of symptoms on function and sports ability in patients with acute hamstring injuries. *Br J Sports Med*. 2014;48:1607–1612.
21. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation. *Value Health*. 2005;8:94–104.
22. Tennant A, Penta M, Tesio L, et al. Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model: the Pro-ESOR project. *Med Care*. 2004;42:37–48.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Krogsgaard MR, Brodersen J, Christensen KB, et al. How to translate and locally adapt a PROM. Assessment of cross-cultural differential item functioning. *Scand J Med Sci Sport*. 2020;00:1–10. https://doi.org/10.1111/sms.13854