International Urogynecology Journal (2019) 30:1989–2000
https://doi.org/10.1007/s00192-019-03920-1

REVIEW ARTICLE

A systematic review of the psychometric properties of the cross-cultural adaptations and translations of the Prolapse Quality of Life (P-QoL) questionnaire

Tadesse Belayneh[1] · Abebaw Gebeyehu[1] · Mulat Adefris[2] · Guri Rortveit[3,4]

Received: 28 November 2018 / Accepted: 5 March 2019 / Published online: 26 April 2019

Abstract

Introduction and hypothesis Due to linguistic and cultural differences, there is a need to test the psychometrics of the translated versions of any patient-reported outcome measures. We investigated the psychometric properties of the Prolapse Quality of Life (P-QoL) questionnaire for non-English-speaking populations by conducting a systematic review of studies that examined the psychometric properties of non-English versions.

Methods We searched PubMed, Latin American and Caribbean Health Science Information Center (LILACS), and Science Direct databases for articles published in English up to February 2018. Methodological quality and quality of psychometric properties were assessed by two independent reviewers using the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) checklist and validated quality assessment criteria. The two assessments were combined to produce the best level of evidence per language/translation.

Results Sixteen articles in 13 languages were retrieved. Most ($n = 9$; 56.3%) were not rigorously translated or reported, and there was poor evidence for structural validity. Internal consistency was reported in all studies, and all studies had good methodological quality. There was fair evidence for construct and good to fair for criterion validity. Evidence for responsiveness was good, although this was evaluated only in three studies.

Conclusions There is limited evidence supporting the psychometric robustness of the original validation and translated versions of P-QoL. Cross-cultural adaptations are insufficient. Given this variability, the individual psychometrics of a translation must be considered prior to use. Responsiveness, measurement error, and cutoff values should also be assessed to increase the clinical utility and psychometric robustness of the translated versions.

Keywords Pelvic organ prolapse · Measurement properties · Assessment scale · Translation

Abbreviations

CFA	Confirmatory factor analysis
COSMIN	Consensus-based Standards for the Selection of Health Measurement Instruments
EFA	Exploratory factor analysis
HRQoL	Health-related quality of life
ICC	Intraclass correlation coefficient
POP	Pelvic organ prolapse
P-QoL	Prolapse Quality of Life questionnaire
SCC	Spearman correlation coefficient

Introduction

Pelvic organ prolapse (POP) occurs when the pelvic floor no longer supports the correct positioning of the pelvic organs [1].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00192-019-03920-1) contains supplementary material, which is available to authorized users.

✉ Tadesse Belayneh
tadbel20@gmail.com

[1] Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, 196, Gondar, Ethiopia

[2] Department of Gynecology and Obstetrics, School of Medicine, University of Gondar, Gondar, Ethiopia

[3] Research Group for General Practice, Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

[4] Research Unit for General Practice, Uni Health, Uni Research, Bergen, Norway

Global prevalence is reported to be ~10% [2], but it is closer to 20% in low-income countries [3]. POP presents with nonspecific symptoms [4, 5] that significantly affect women's health-related quality of life (HRQoL) [3, 6]. In recent years, there has been growing interest in incorporating patient-reported outcomes in clinical research [7]. HRQoL is a central patient-reported outcome and has become more widely accepted as a valid goal within patient treatment [8]. As such, it is used in management of POP [9, 10] to evaluate the effectiveness of different treatment modalities [11], to tailor patient care, and to improve satisfaction with care [12–14]. Given the growing use of HRQoL as an outcome measure [14], considerable effort has been put into development and translation of HRQoL questionnaires in the field of urogynecology [9]. A review by Al-Badr [11], for instance, identified four instruments designed to evaluate HRQoL in women with POP, including the Pelvic Floor Distress Inventory (PFDI; 46 items), the Pelvic Floor Impact Questionnaire (PFIQ; 93 items) [15], the Prolapse Quality of Life questionnaire (P-QoL; 20 items) [16], and the electronic Personal Assessment Questionnaire Pelvic Floor (ePAQ-PF; 58 items) [17]. Compared with other instruments, P-QoL has a relatively small number of items and covers nine domains of HRQoL (general health, prolapse impact, role limitations, physical limitations, social limitations, personal relationships, emotional problems, sleep/energy disturbance, and severity of symptoms) [16].

In modern health-care systems, patient-centered care and evidence-based practice have led researchers to develop, adapt, and translate different outcome measures [14, 18]. To this end, the P-QoL has undergone extensive translations in several languages [19–23]. However, evidence for the psychometric properties of the translated versions is fragmented [24]. Given the wide variation in P-QoL translations due to linguistic and cultural variations, a systematic evaluation of the translated versions will assist in providing information on the quality of the prevailing tools, as well as gaps in knowledge and adaptation. Further, because the reported psychometrics are dependent on the quality of the methodologies applied, methodologies of available studies need to be appraised in accordance with the Consensus-based Standards for the Selection of Health Status Measurement Instruments (COSMIN) criteria [25, 26]. Therefore, the objectives of our review were to: (I) systematically identify the translated non-English versions of the P-QoL, (II) evaluate the methodologies implemented in the adaptation process, and (III) appraise the psychometric properties of the translated and adapted versions.

Materials and methods

Protocol and registration

First, the Database of Abstracts of Reviews of Effects (DARE) (http://www.library.UCSF.edu) was explored to determine whether any systematic reviews exist and to check the availability of ongoing projects associated with the subject. Then, the protocol for this review was prospectively registered with the PROSPERO database of systematic reviews (CRD42018084601) (https://www.crd.york.ac.uk/prospero/#myprospero).

Data sources and search strategy

We searched PubMed, Latin American and Caribbean Health Sciences Literature (LILACS), and Science Direct (Supplementary S1) databases. Google and Google Scholar were additionally searched to identify gray literature. Reference lists of the identified articles were manually searched for possible articles for inclusion in order to reach literature saturation [26]. The search strategy covered the following four aspects, as per Terwee et al.'s [27] recommendation: (1) condition of interest: pelvic organ prolapse, (2) patient-reported outcome measures, (3) measurement properties, and (4) exclusion filter. The search terms were used separately and in combination using Boolean operators like OR or AND. All retrieved articles collected from the electronic and manual searches were entered into the EndNote-7 software (Thomson Reuters, New York, NY, USA).

Eligibility criteria

All English-language articles published before 20 February 2018, were considered eligible. Studies were included if one of the aims was translation of the P-QoL and evaluation of at least one measurement property (reliability, validity, and/or responsiveness) of the translated version. Whenever available, studies were also included if they provided additional psychometric properties of the instrument that were not evaluated during the translation.

Data extraction

All relevant characteristics of the included studies were extracted (TB), and completeness was independently verified (MA). Participant age, country in which the study was conducted, method used to select participants, sample size in which the instrument was tested, settings in which the study was conducted, and language in which the instrument was evaluated and analyzed were extracted using a preconstructed extraction form. Furthermore, for each translated version, methodological quality of the translation process and evidence for psychometric properties were extracted. The following domains of measurement properties were extracted: (1) reliability [internal consistency (IC) and test–retest reliability], (2) validity (content, construct, structural, cross-cultural, and criterion validity), and (3) responsiveness [26].

Assessment of methodological quality

Methodological quality was appraised using the COSMIN checklist [28], which comprises nine boxes, each dealing with one defined measurement property: Internal consistency, reliability, measurement error, content validity, structural validity, hypotheses testing, and cross-cultural validity (these three are aspects of construct validity), criterion validity, and responsiveness [26, 29]. Each item is rated on a 4-point scale (poor, fair, good, excellent), and an overall score is determined by a worst-score-counts procedure [25, 26, 30]. Two reviewers (TB and MA) independently assessed the quality. They met for a calibration review in which they independently reviewed one study, then met and discussed each item of the COSMIN list to clarify its meaning and interpretation. Following this, the methodological quality was rated across a set of items related to each attribute independently by the same reviewers. In cases of disagreement, a third reviewer (AG) performed a separate appraisal, following which consensus was reached in each case.

Best-evidence synthesis

To come up with best-evidence synthesis for each psychometric property, we used Terwee et al.'s criteria [31] (Table 1) and the COSMIN methodological assessment checklist [25, 26, 30] (Table 2). Each measurement property per study was rated as either positive (+), indeterminate (?), negative (−), or no information (0). For example, structural validity was rated as positive if the factors identified after performing a factor analysis were reported to explain at least 50% of the variance. If the factors explained <50% of variance, structural validity was rated as negative. The indeterminate rating was given if the percentage of variance explained was not reported. The level of evidence for each measurement property is outlined in Table 3.

Results

Among the 2771 studies screened, 95 were selected based on titles and abstracts. After reading the full text and applying selection criteria, 16 studies were analyzed for review [16, 19–22, 32–42] (Fig. 1). One study was the original questionnaire [16] and the remaining were translated versions

Participants and settings

A consecutive sample was the most common method of participant selection ($n = 6$; 37.5%), and most used a longitudinal design ($n = 12$, 75.0%). Almost all were either from middle- or high-income settings ($n = 15$; 93.7%), and most were conducted at university gynecologic clinics ($n = 12$; 75.0%). Number of participants varied greatly across studies, ranging from 39 to 292. A summary of articles is presented in Table 4.

Adaptations and translations

For all translations ($n = 15$), response options were similar to the original 4-point Likert scale, and the direction of the response was described. None of the original terms on the P-QoL were modified or changed, and all included 20 items and nine domains. However, most studies did not use a translation guide ($n = 11$, 73.3%). The P-QoL was self-administered in most studies ($n = 11$, 68.7%) (Table 4).

Individual translated studies

Detailed description of study psychometric properties are shown in Supplementary S2 and S3.

Afrikaans

This version was described in one study [19]. The methodology for cross-cultural validation was poor. Only solitary forward translation was performed, reconciliation of translations was poorly described, and the tool was not reviewed by a committee. The evidence for IC was limited because a moderate sample size was used ($n = 39$), and three of seven domains yielded a Cronbach's alpha <0.70. Evidence for test–retest reliability was unknown. Stability for the retest sample was not clearly outlined, and test conditions were not similar. There was unknown evidence for structural validity because neither exploratory factor analysis (EFA) nor confirmatory factor analysis (CFA) were performed. There was limited evidence for construct validity because no specific hypotheses were formulated (despite it being possible to deduce what was expected) and the sample size was suboptimal ($n = 39$). Although it was unclear whether the criterion used can be considered an adequate gold standard, there was moderate positive evidence for criterion validity.

Chinese (traditional version)

One study was available [42]. The methodology for cross-cultural validation was poor. Solitary backward translation was performed, but whether the translations were done independently was not clear, and the tool was not reviewed by a committee. Evidence for IC and test–retest reliability was poor. The IC statistic was not calculated for each subscale separately, stability for the retest sample ($n = 30$) was not clearly outlined, and Spearman's correlation coefficient (SCC) was less than the recommended value in six of the nine domains. There was limited evidence for construct validity because the minimum number of hypotheses formulated a priori was not confirmed. Evidence for criterion validity was

Table 1 Terwee et al.'s quality criteria for measurement properties

Measurement property	Definition	Rating with quality criteria [a,b]
Content validity (including face validity)	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire	+ A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection AND target population and (investigators OR experts) were involved in item selection ? A clear description of above-mentioned aspects is lacking OR only target population involved OR doubtful design or method _ No target population involvement; criteria for + not met 0 No information found on target population involvement
Construct validity	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	+ 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); no correlations with instrument(s) measuring related construct(s) AND no differences between relevant groups reported _ Less than 75% of hypotheses were confirmed, despite adequate design and methods 0 No information found on construct validity
Criterion validity	The extent to which scores on a particular questionnaire relate to a gold standard	+ Convincing arguments that gold standard is gold AND correlation with gold standard ≥0.70 ? No convincing arguments that gold standard is gold OR doubtful design or method _ Correlation with gold standard <0.70 despite adequate design and method 0 No information found on criterion validity.
Internal consistency	The extent to which items in a (sub)scale are inter-correlated, thus measuring the same construct	+ Factor analyses performed on adequate sample size (7 × no. 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) <0.70 or > 0.95, despite adequate design and method 0 No information found on internal consistency
Agreement (reproducibility)	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	+ MIC < 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
Reliability (reproducibility)	The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)	+ ICC or weighted kappa >0.70 ? Doubtful design or method (e.g., time interval not mentioned) _ ICC or weighted kappa <0.70, despite adequate design and method 0 No information found on reliability
Responsiveness	The ability of a questionnaire to detect clinically important changes over time	+ SDC or SDC < MIC OR MIC outside the LOA OR RR >1.96 OR AUC ≥0.70 ? Doubtful design or method _ SDC or SDC > MIC OR MIC equals or inside LOA OR RR ≤1.96 OR AUC < 0.70, despite adequate design and methods; 0 No information found on responsiveness.
Floor and ceiling effects	The number of respondents who achieved the lowest or highest possible score	+ ≤15% of the respondents achieved the highest or lowest possible scores ? Doubtful design or method _ >15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods 0 No information found on interpretation
Interpretability	The degree to which one can assign qualitative meaning to quantitative scores	+ Mean and SD scores presented of at least four relevant subgroups of patients and MIC defined ? Doubtful design or method OR less than four subgroups OR no MIC defined 0 No information found on interpretation

MIC minimal important change, *SDC* smallest detectable change, *LOA* limits of agreement, *ICC* Intraclass correlation, *SD* standard deviation, *AUC* area under the curve, *RR* responsiveness ratio

[a] + positive rating, ? indeterminate rating, _ negative rating, 0 no information available

[b] Doubtful design or method lacking of a clear description of the design or methods of the study, sample size smaller than 50 subjects (should be at least 50 in every (subgroup) analysis), or any important methodological weakness in the design or execution of the study

Table 2 Methodological quality rating per measurement property

Authors/year	Version – country	Internal consistency	Test–retest	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsiveness
Brandt et al. 2016	Afrikaans – South Africa	Fair	Poor	Poor	Fair	Poor	Fair	ND
Chuang et al. 2016	Traditional Chinese – Taiwan	Poor	Poor	Poor	Fair	Poor	Fair	ND
Claerhout, F., et al. 2010	Dutch – Belgium	Excellent	Poor	Poor	Good	Poor	Good	Fair
Veit-Rubin, N., et al. 2015	French – France	Fair	Poor	Poor	Fair	Poor	Fair	ND
Lenz, F., et al. 2009	German – Germany	ND	ND	Fair	Fair	Poor	Fair	ND
Digesu, G. A., et al. 2003	Italian – Italy	Good	Fair	Fair	Good	Poor	Fair	ND
Nojomi, M., et al. 2012	Persian – Iran	Excellent	Fair	Poor	Fair	Fair	Good	ND
Morovatdar, N., et al. 2015	Persian – Iran	ND	ND	ND	ND	ND	ND	Good
Rzepka, J., et al. 2016	Polish – Poland	Excellent	ND	Poor	Fair	Poor	Good	ND
Scarlato, A., et al. 2011	Portuguese – Brazil	Good	Poor	Poor	Fair	Fair	Good	Good
de Oliveira, M., et al. 2009	Portuguese – Brazil	Good	Poor	Poor	Good	Excellent	Good	ND
Svihrova, V., et al. 2010	Slovakian – Slovakia	Excellent	Poor	Poor	Good	Poor	Good	ND
Flores-E, C., et al. 2015	Spanish – Chile	Good	ND	Poor	Good	Excellent	Good	ND
Manchana, T. 2010	Thai – Thailand	Excellent	Poor	Poor	Fair	Poor	Fair	ND
Cam, C., et al. 2007	Turkish – Turkey	Poor	Poor	Poor	Fair	Fair	Fair	ND
Digesu, G. A., et al. 2005	English – UK	Excellent	Poor	Poor	Excellent	ND	Excellent	ND

ND Not determined

inconclusive because psychometrics of the purported gold standard measure were questionable and domain scores were less than the recommended value.

Dutch

One study was retrieved [36]. The methodology for cross-cultural validation was poor. There was no description as to whether translations were done independently, and only solitary forward–backward translation was performed. There was strong evidence for IC and hypothesis testing, but evidence for test–retest reliability was poor because test conditions and stability of the retest sample were not clearly outlined. Further, a small sample was used (n = 20). There was indeterminate evidence for structural validity because factor analysis was not performed, and evidence for criterion validity was conflicting. There was no correlation between stages of prolapse and P-QoL for symptomatic women. Responsiveness was rated fair because hypotheses were vague, but it was possible to deduce what was expected.

French

One study was available [22]. The methodology for cross-cultural validation was poor. The tool was not pretested in the target population, and only solitary forward–backward translations were performed. There was moderate evidence for IC because of the moderate sample size ($n = 47$). Evidence for test–retest reliability was poor. Test conditions and stability of the retest sample were not clearly outlined, and a suboptimal sample size was used ($n = 25$). There was indeterminate evidence for structural validity because factor analysis was not performed and a suboptimal sample was utilized. Evidence for construct and criterion validity was fair. Sample size was moderate, hypotheses were vague, and it was unclear whether criteria could be considered an adequate gold standard.

German

We found one study [35]. The methodology for cross-cultural validation was poor. The reconciliation process of the

Table 3 Synthesis of results for each measurement property

Authors/year	Version – country	Internal consistency	Test–retest	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsiveness
Brandt, C., et al. 2016	Afrikaans – South Africa	+	?	?	+	?	++	0
Chuang, F. C., et al. 2016	Chinese – Taiwan	?	?	?	+	?	±	0
Claerhout, F., et al. 2010	Dutch – Belgium	+++	?	?	++	?	±	+
Veit-Rubin, N., et al. 2015	French – France	++	?	?	+	?	+	0
Lenz, F., et al. 2009	German – Germany	0	0	?	+	?	+	0
Digesu, G. A., et al. 2003	Italian – Italy	++	++	?	+	?	±	0
Nojomi, M., et al. 2012	Persian – Iran	+++	++	?	++	++	+	0
Morovatdar, N., et al. 2015	Persian – Iran	0	0	0	0	0	0	++
Rzepka, J., et al. 2016	Polish – Poland	+++	0	?	+	?	±	0
Scarlato, A., et al. 2011	Portuguese – Brazil	++	?	?	+	++	+	++
de Oliveira, M., et al. 2009	Portuguese – Brazil	++	?	?	++	+++	+	0
Svihrova, V., et al. 2010	Slovakian – Slovakia	+++	?	?	+	?	+	0
Flores-E, C., et al. 2015	Spanish – Chile	++	0	?	++	+++	±	0
Manchana, T. 2010	Thai – Thailand	+++	?	?	++	?	+	0
Cam, C., et al. 2007	Turkish – Turkey	?	?	?	++	++	+	0
Digesu, G. A., et al. 2005	English – UK	+++	?	?	++	0	+	0

+++ strong positive evidence, ++ moderate positive evidence, + limited positive evidence, ± conflicting findings, ? unknown due to poor methodological quality, 0 not determined

translation was poorly described, there was no description of whether translations were done independently, and the tool was not reviewed by a committeeor pre-tested; only solitary forward–backward translations were performed, and there was a lack of a harmonized final version. Because IC and test–retest reliability were not evaluated, evidence was not generated. There was fair evidence for construct, criterion, and structural validity, although factor analysis was not performed. Specific hypotheses were formulated and confirmed.

Italian

One study was found [32]. The methodology for cross-cultural validation was poor. Backward translation was not performed or was poorly described, and the tool was not pretested. There was good evidence for IC. Evidence for test–retest reliability was fair because stability for sample was not clearly outlined. Evidence for criterion validity was conflicting because the SCC was less than the recommended value in almost all domains. There was moderate evidence for construct validity because specific hypotheses were formulated. There was limited evidence for structural validity because EFA and CFA were not performed.

Persian

Two studies were available [40, 41]. The first aimed to translate and evaluate reliability and validity, and the second aimed to evaluate responsiveness. Methodology for cross-cultural validation was fair. Only solitary forward–backward translation was performed; reconciliation of translations was not outlined, and the tool was not reviewed by a committee. There was strong evidence for IC and moderate evidence for test–retest reliability. Reliability test utilized a suboptimal sample size ($n = 30$). Evidence for criterion validity was good, although psychometrics of the purported gold standard measure were questionable. There was poor evidence for structural validity because EFA or CFA was not performed [40]. There was strong evidence for responsiveness of the tool [41].

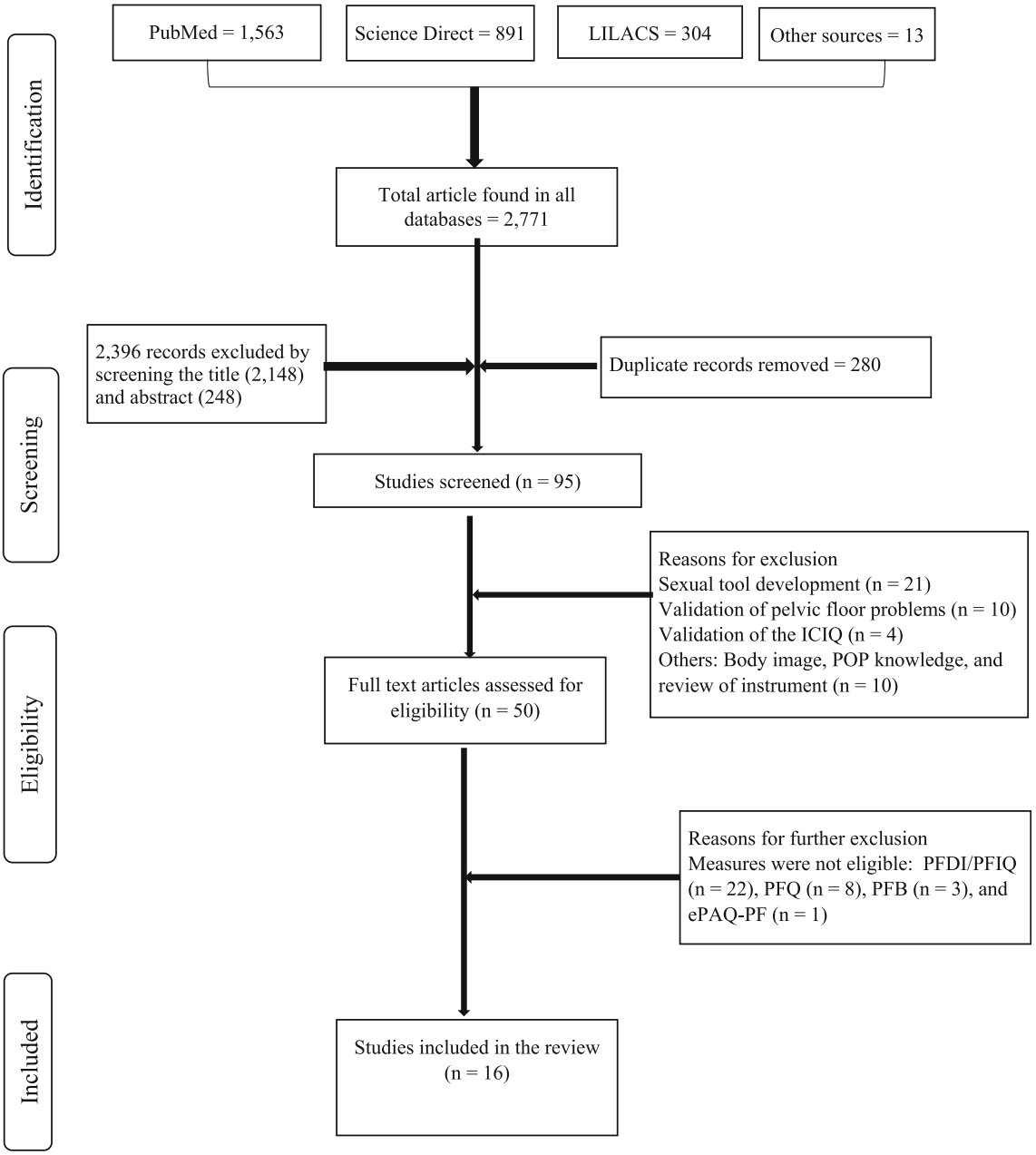

Fig. 1 Systematic review methodology in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines

Polish

A single study was available [21]. The tool was poorly translated because only solitary backward translation was performed, reconciliation of the translation was not outlined, and the tool was neither reviewed by a committee nor pretested. There was strong evidence for IC, but test–retest reliability was not performed. Evidence for criterion validity was good, although psychometrics of the purported gold standard measure were questionable. There was poor evidence for structural validity because EFA or CFA was not performed.

Portuguese

Brazilian version The methodology for cross-cultural validation was fair. The tool was not reviewed by a committee, and only solitary backward translation was performed. There was moderate evidence for IC, but poor evidence for test–retest reliability. Stability for the retest sample (unknown sample size) was not clearly outlined, and there was a disparity in questionnaire administration. There was poor evidence for structural validity because EFA or CFA was not performed. There was unknown evidence for construct validity because no specific hypotheses were formulated and sample size was

Table 4 Study characteristics

Author/year	No. participants	Age [years; mean (SD) or range]	Country	Language	Study aim(s)	Setting	Sampling	Mode of delivery	Design	Analyses
Digesu, G. A., et al. (2005)	215 (140[a], 75[b])	53.1 (19–83)[a], 53.1 (32–79)[b]	United Kingdom	English	PT	UGC	Conv	SA	CS	NPT
Digesu, G. A., et al. (2003)	184 (126[a], 58[b])	54 (33–82)[a], 41 (19–83)[b]	Italy	Italian	T, PT	UGC	Conv	SA	L	NPT
Brandt, C., et al. (2016)	39 (25[a], 14[b])	Median 60[a], 45.5[b]	South Africa	Afrikaans	T, PT	UGC	Conv	SA	L	NPT
Cam, C., et al. (2007)	292 (145[a], 147[b])	44.2 ± 9.8[a], 37.0 ± 9.8[b]	Turkey	Turkish	T, PT	UGC	NS	M	L	NPT
Chuang, F. C., et al. (2016)	244 (159[a], 85[b])	62 (35–86)[a], 61 (33–84)[b]	Taiwan	Traditional Chinese	T, PT	UroGC	Conv	M	L	NPT
Claerhout, F., et al. (2010)	228 (155[a], 73[b])	65 ± 10[a], 57 ± 7.8[b]	Belgium	Dutch	T, PT	UPF[a], GC[b]	P	SA	L	NPT
de Oliveira, M. S., et al. (2009)	98 (68[a], 30[b])	57.4 ± 13.1[a], 47.9 ± 12.3[b]	Brazil	Portuguese	T, PT	UroGC	P	M	L	NPT
Flores C., et al. (2015)	128[a]	53.1 ± 9.8[a]	Chile	Spanish	T, PT	UGC	NS	SA	CS	ANOVA
Lenz, F., et al. (2009)	215 (140[a], 75[b])	64 (30–87)[a], 58 (26–101)[b]	Germany	German	T, PT	UGC	P	SA	CS	NPT
Manchana, T. et al.; (2010)	175 (115[a], 60[b])	Median 67 (32–85)[a], 53.5(29–73)[b]	Thailand	Thai	T, PT	UroGC	Cons	M	L	NPT
Nojomi, M., et al. (2012)	249 (124[a], 125[b])	44.9[a], 35.7[b]	Iran	Persian	T, PT	UGC	Cons	SA	L	NPT
Morovatdar, N., et al. (2015)	50[a]	43.5 ± 11.8[a]	Iran	Persian	PT	UGC	P	SA	L	NPT
Rzepka, J., et al. (2016)	231 (154[a], 77[b])	65.6 (33–91)[a], 64.9 (30–85)[b]	Poland	Polish	T, PT	UGC	Cons	SA	CS	NPT
Scarlato, A., et al. (2011)	65 (45[a], 20[b])	58.3 ± 13[a], 63.8 ± 9.7[b]	Brazil	Brazilian Portuguese	T, PT	UGC	Cons	IA	L	NPT
Svihrova, V., et al. (2010)	129 (50[a], 79[b])	45[a], 37[b]	Slovakia	Slovakian	T, PT	UGC	Cons	SA	L	NPT
Veit-Rubin, N., et al. (2015)	47 (25[a], 22[b])	Median 61 (39–81)[a], 35 (16–76)[b]	France	French	T, PT	UGC	Cons	SA	L	NPT

T translation, *PT* psychometric testing, *SD* standard deviation, *EFA* exploratory factor analysis, *CFA* confirmatory factor analysis, *NPT* nonparametric test, *UGC* university gynecology clinic, *UPF* university pelvic floor, *UroGC* university urogynecology clinic, *CS* cross-sectional, *L* longitudinal, *SA* self-administered, *IA* interview administer, *M* mixed, *Conv* convenience, *P* purposive, *Cons* consecutive, *NS* not stated

[a] Symptomatic

[b] Asymptomatic

small ($n = 65$). Evidence for responsiveness was considered moderate because a suboptimal sample was used ($n = 56$) [39].

Portuguese version Methodology for cross-cultural validation was good, although reflection of the original version was inadequate. There was moderate evidence for IC, unknown evidence for test–retest reliability because the setting and number of participants were not clearly mentioned, and it was doubtful whether measurements were independent. There was poor evidence for structural validity because EFA or CFA was not performed. There was limited evidence for construct validity due to the small sample size ($n = 98$). Specific hypotheses were formulated [34].

Slovakian

One study was available [38]. The methodology for cross-cultural validation was poor. The tool was not pretested, and only solitary backward translation was performed. There was good evidence for IC, but evidence for test–retest reliability was poor. Stability, setting, and condition of administration for the retest sample were not clearly outlined. There was poor evidence for structural validity because EFA or CFA was not performed. There was good evidence for construct and criterion validity despite specific hypotheses not being formulated and psychometrics of the purported gold standard measure being questionable.

Spanish

One study was retrieved [20]. There was strong evidence for cross-cultural validation, moderate evidence for IC, and poor evidence for structural validity because EFA or CFA was not performed. There was good evidence for construct and criterion validity despite psychometrics of the purported gold standard measure being questionable.

Thai

A single study was found [37]. The methodology for cross-cultural validation was poor. Only solitary forward–backward translation was performed, reconciliation of the translation was not outlined, and the tool was neither reviewed by a committee nor pretested. There was good evidence for IC, but evidence for test–retest reliability was poor. Stability and condition of administration for the retest sample ($n = 20$) were not clearly outlined. There was poor evidence for structural validity because EFA or CFA was not performed. There was fair evidence for construct and criterion validity despite psychometrics of the purported gold standard measure being questionable.

Turkish

A single study was found [33]. The methodology for cross-cultural validation was good. Reconciliation of translation was not outlined, and only solitary backward translation was performed. There was indeterminate evidence for IC and test–retest reliability. Cronbach's alpha was not reported. Stability and condition of administration for the retest sample ($n = 30$) were not clearly outlined. There was poor evidence for structural validity because neither EFA nor CFA was performed. There was fair evidence for construct and criterion validity despite the psychometrics of the purported gold standard measure being questionable.

Discussion

To provide scientifically robust results and to establish scientific integrity and clinical utility, psychometric properties of a newly developed or translated version of an instrument have to be evaluated [43]. This is the first systematic review that evaluates measurement properties of translations of the P-QoL questionnaire using the COSMIN checklist.

Quality of cross-cultural validity/translations

Transcultural adaptation, translation, and validation aim to succinctly capture the meaning of latent constructs in another population [18]. As such, a rigorous translation process is essential [44]. Cross-cultural validity was examined in 14 studies, of which only two were translated using robust methodologies [20, 34] and nine were of poor methodological quality [19, 21, 22, 32, 35–38, 42]. The lack and/or scarceness of descriptions of whether the translations were done independently, the poor reconciliation processes of the translations, the presence of only a solitary forward and/or backward translation, the absence of a panel of experts, and the lack of pretesting of the translated version in the target population compromised the methodological rigor of most reviewed studies. These methodological limitations could affect the ability to produce a culturally acceptable translation [18, 30]. Because each population has its own unique perspectives in regard to HRQoL, and because the lack of quality of translations affects generalizability and comparability of study findings, authors of such studies should utilize rigorous methodologies for future translations and perform further validation to produce a culturally sensitive instrument.

Reliability

Internal consistency was examined in 13 studies, of which six scored excellent [16, 21, 36–38] and four scored good [20, 32, 34, 39]. In most studies, questionnaire displayed adequate

evidence with Cronbach's alpha values ≥0.7 [16, 20, 21, 32, 34, 36–40]. Lower-quality ratings were assigned mostly due to the absence of IC for every subscale of the P-QoL questionnaire [42] or to an inadequate number of study participants [19, 22, 33]. Given the limitations in the structural validity testing and lack of rigor in the translation process, results for IC might need to be interpreted with caution. This is because a tool can be reliable yet not be valid [18, 45].

Test–retest reliability

This provides evidence for the stability of a questionnaire over time [30] and was assessed in 11 of 15 studies. However, most studies, except for one using ICC [36], reported the correlation between the paired test–retest reliability by SCC [16, 22, 32–34, 37, 38, 42], which is not in agreement with the COSMIN criteria. The majority of studies scored poor for test–retest reliability [16, 19, 22, 33, 34, 36–39, 42] largely due to the small sample sizes, poor description of test conditions, and poor stability of the retest sample. A 2-week test–retest reliability was performed in all studies, except for the French [22], Italian [32], and Dutch [36] studies.

Measurement error

This difficult to assess because the studies provided no information on parameters such as minimal important change, standard error of measurement, or limits of agreement. Likewise, none reported the use of modern psychometric methods (item response theory or Rasch measurement theory). Thus, we did not assess study quality for this domain.

Construct validity

Hypothesis testing

This was reported in eight of the 15 studies either as clearly formulated hypotheses with the magnitudes and directions of correlations or as the possibility of deducing what was expected if there was an absence of such prior specific hypotheses. For this reason, most studies were rated as fair for the methodological quality of their hypothesis testing. The absence of a clear hypothesis makes it difficult to determine whether any results reported for construct validity were due to chance or not [28].

Structural validity

Structural validity, as assessed by factor analysis, gives evidence as to whether all items in a questionnaire can be summed up into a single overall score or whether different domains need to be scored separately [26]. It is usually performed using EFA and/or CFA: EFA is used to identify underlying factors, whereas CFA is used to verify the factor structure [46, 47]. It is important that studies perform factor analysis because it verifies the scale structure that determines scoring and interpretation of a measure's internal consistency statistic [48]. P-QoL is a multidimensional and subjective construct that is dependent on sociocultural contextual factors [16]. Although not all studies assessing structural validity found the same factor structure, we found a similar factor structure as postulated by the developers of the P-QoL [18]. In this review, there were no studies on structural validity and there is no evidence for structural validity.

Criterion validity

This Is defined as the extent to which scores on an outcome measure perform against an established gold standard [31]. This was not assessed for any study because the COSMIN Delphi panel does not regard any patient-reported outcome measure to be a true gold standard [26]. Given that P-QoL is a construct variable [16], it is difficult to establish a gold standard against which the P-QoL can be assessed. In studies in our review, all used prolapse stage scores as the gold standard. Nevertheless, for the 15 studies that evaluated criterion validity, psychometrics of the purported gold standard were either questionable or poorly described. Therefore, evidence for criterion validity was rated as good to fair.

Responsiveness

This is an important characteristic of a questionnaire, especially when used to measure a change as a result of intervention. However, only three studies reported on P-QoL responsiveness. There was good evidence found for responsiveness in each study.

Strengths and weaknesses

A strength of this review is the comprehensive independent search strategy limiting the chances of missing important studies. Another strength is use of the standardized COSMIN methodology for critical appraisal of methodological quality, quality of the questionnaires, and the level of evidence. However, there are also limitations to this review. First, the use of the COSMIN checklist to evajuate methodological quality might have been a potential limitation. This is because the checklist came into effect in 2011. and half the translations had been performed prior to its publication. Second, we included studies only from peer-reviewed English-language journals, which might have restricted our findings. The P-QoL questionnaire was assessed in a large variety of languages, but whether results for individual questionnaires discussed in this review can be validly generalized across language versions is not clear. It is clear that questionnaires

may perform differently across different languages and cultures [48]. Third, the raters might have had to make a large number of judgments for each psychometric property. Although the COSMIN checklist [28] and quality criteria for measurement properties [31] being defined as objectively as possible, different raters might come to different conclusions. To reduce this risk, in case of disagreement between the two reviewers, a third reviewer was consulted.

In summary, this systematic review adds to the current literature by providing a structured and comprehensive overview of the measurement properties and the methodological quality of translations of the P-QoL. Specifically, this review is of use when translating and adapting the questionnaire in the future. Many of the translations did not follow a rigorous translation process and had poor evidence for structural validity, thus caution is advised in the use of these questionnaires. In addition, future translation studies must perform independent multiple backward–forward translations, reconcile translations, use a panel of experts, and perform factorial analysis to increase clinical utility and psychometric robustness of translated versions. Because not all measurement properties (measurement error and responsiveness) of the P-QoL have been extensively evaluated, future validation studies should also emphasize these psychometric properties.

Acknowledgements We thank Dr. Tadesse Awoke and Dr. Mekuriaw Alemayehu for their assistance in developing the search strategy for the electronic databases used in this systematic review.

Compliance with ethical standards

Conflicts of interest None.

References

1. Haylen BT, Maher CF, Barber MD, Camargo S, Dandolu V, Digesu A, et al. An International Urogynecological Association (IUGA) / International Continence Society (ICS) joint report on the terminology for female pelvic organ prolapse (POP). Neurourol Urodyn. 2016;35:137–68. https://doi.org/10.1002/nau.22922.
2. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. Lancet. 2012;380(9859):2163–96.
3. Walker GJ, Gunasekera P. Pelvic organ prolapse and incontinence in developing countries: review of prevalence and risk factors. Int Urogynecol J. 2011;22:127–35. https://doi.org/10.1007/s00192-010-1215-0.
4. Machin SE, Mukhopadhyay S. Pelvic organ prolapse: review of the aetiology, presentation, diagnosis and management. Menopause Int. 2011;17(4):132–6. https://doi.org/10.1258/mi.2011.011108.
5. Maher CM, Feiner B, Baessler K, Glazener CM. Surgical management of pelvic organ prolapse in women: the updated summary version Cochrane review. Int Urogynecol J. 2011;22:1445–57. https://doi.org/10.1007/s00192-011-1542-9.
6. Chaudhary A. Health related quality of life of women suffering from Uterine Prolapse before and after six months of surgery. Health. 2014;6:350–5. https://doi.org/10.4236/health.2014.65051.
7. Speight J, Barendse SM. FDA guidance on patient reported outcomes. BMJ: British Medical Journal (Online). 2010;340:c2921.
8. Sprangers MA. Disregarding clinical trial-based patient-reported outcomes is unwarranted: five advances to substantiate the scientific stringency of quality-of-life measurement. Acta Oncol. 2010;49(2):155–63.
9. Doaee M, Moradi-Lakeh M, Nourmohammadi A, Razavi-Ratki SK, Nojomi M. Management of pelvic organ prolapse and quality of life: a systematic review and meta-analysis. Int Urogynecol J. 2014;25:153–63. https://doi.org/10.1007/s00192-013-2141-8.
10. Reid F. Assessment of pelvic organ prolapse: a practical guide to the pelvic organ prolapse quantification. Obstet Gynaecol Reprod Med. 2014;24:170–6. https://doi.org/10.1016/j.ogrm.2014.04.011.
11. Al-Badr A. Quality of Life Questionnaires for the Assessment of Pelvic Organ Prolapse: Use in Clinical Practice. Lower Urinary Tract Symptoms. 2013;5(3):121–8. https://doi.org/10.1111/luts.12006.
12. Barber MD, Brubaker L, Nygaard I, Wheeler TL 2nd, Schaffer J, Chen Z, et al. Defining success after surgery for pelvic organ prolapse. Obstet Gynecol. 2009;114(3):600–9. https://doi.org/10.1097/AOG.0b013e3181b2b1ae.
13. Lee U, Raz S. Words of wisdom. Re: defining success after surgery for pelvic organ prolapse. Eur Urol. 2010;58(4):633–4.
14. Toozs-Hobson P, Freeman R, Barber M, Maher C, Haylen B, Athanasiou S, et al. An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for reporting outcomes of surgical procedures for pelvic organ prolapse. Neurourol Urodyn. 2012;31(4):415–21.
15. Barber MD, Kuchibhatla MN, Pieper CF, Bump RC. Psychometric evaluation of 2 comprehensive condition-specific quality of life instruments for women with pelvic floor disorders. Am J Obstet Gynecol. 2001;185(6):1388–95. https://doi.org/10.1067/mob.2001.118659.
16. Digesu GA, Khullar V, Cardozo L, Robinson D, Salvatore S. P-QOL: a validated questionnaire to assess the symptoms and quality of life of women with urogenital prolapse. Int Urogynecol J. 2005;16:176–81. https://doi.org/10.1007/s00192-004-1225-x.
17. Jones G, Radley S, Lumb J, Farkas A. Responsiveness of the electronic personal assessment questionnaire-pelvic floor (ePAQ-PF). Int Urogynecol J. 2009;20(5):557–64.
18. 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.
19. Brandt C, Van Rooyen C, Cronje H. Validation of the prolapse quality of life questionnaire (PQOL): an Afrikaans version in a South African population. S Afr J Obstet Gynaecol. 2016;22(2):38.
20. Flores-Espinoza C, Araya AX, Pizarro-Berdichevsky J, Santos V, Ferrer M, Garin O, et al. Validation of the Spanish-language version of the prolapse quality of life questionnaire in Chilean women. Int Urogynecol J. 2015;26(1):123–30. https://doi.org/10.1007/s00192-014-2484-9.
21. Rzepka J, Zalewski K, Stefanowicz A, Khullar V, Swift S, Digesu GA. Validation of the Polish version of P-QoL questionnaire. Ginekol Pol. 2016;87:477–83.
22. Veit-Rubin N, Digesu A, Swift S, Khullar V, Kaelin Gambirasio I, Dallenbach P, et al. Validation of the French version of the P-QoL questionnaire. Eur J Obstet Gynecol Reprod Biol. 2015;192:10–6. https://doi.org/10.1016/j.ejogrb.2015.05.028.
23. Belayneh T, Gebeyehu A, Adefris M, Rortveit G, Genet T. Translation, transcultural adaptation, reliability and validation of the pelvic organ prolapse quality of life (P-QoL) in Amharic. Health Qual Life Outcomes. 2019;17(1):12.

24. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. 2010;68(4):319–23.
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(4):651–7. https://doi.org/10.1007/s11136-011-9960-1.
26. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–45.
27. 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(8):1115–23.
28. 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(4):539–49.
29. Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol. 2010;10(1):82.
30. Prinsen C, Mokkink L, Bouter L, Alonso J, Patrick D, de Vet H, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147–57.
31. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.
32. Digesu GA, Santamato S, Khullar V, Santillo V, Digesu A, Cormio G, et al. Validation of an Italian version of the prolapse quality of life questionnaire. Eur J Obstet Gynecol Reprod Biol. 2003;106(2):184–92.
33. Cam C, Sakalli M, Ay P, Aran T, Cam M, Karateke A. Validation of the prolapse quality of life questionnaire (P-QOL) in a Turkish population. Eur J Obstet Gynecol Reprod Biol. 2007;135(1):132–5. https://doi.org/10.1016/j.ejogrb.2007.06.009.
34. de Oliveira MS, Tamanini JT, de Aguiar Cavalcanti G. Validation of the prolapse quality-of-life questionnaire (P-QoL) in Portuguese version in Brazilian women. Int Urogynecol J Pelvic Floor Dysfunct. 2009;20(10):1191–202. https://doi.org/10.1007/s00192-009-0934-6.
35. Lenz F, Stammer H, Brocker K, Rak M, Scherg H, Sohn C. Validation of a German version of the P-QOL questionnaire. Int Urogynecol J Pelvic Floor Dysfunct. 2009;20(6):641–9. https://doi.org/10.1007/s00192-009-0809-x.
36. Claerhout F, Moons P, Ghesquiere S, Verguts J, De Ridder D, Deprest J. Validity, reliability and responsiveness of a Dutch version of the prolapse quality-of-life (P-QoL) questionnaire. Int Urogynecol J. 2010;21(5):569–78. https://doi.org/10.1007/s00192-009-1081-9.
37. Manchana T, Bunyavejchevin S. Validation of the prolapse quality of life (P-QOL) questionnaire in Thai version. Int Urogynecol J. 2010;21(8):985–93. https://doi.org/10.1007/s00192-010-1107-3.
38. Svihrova V, Digesu GA, Svihra J, Hudeckova H, Kliment J, Swift S. Validation of the Slovakian version of the P-QOL questionnaire. Int Urogynecol J. 2010;21(1):53–61. https://doi.org/10.1007/s00192-009-0989-4.
39. Scarlato A, Souza CC, Fonseca ES, Sartori MG, Girao MJ, Castro RA. Validation, reliability, and responsiveness of prolapse quality of life questionnaire (P-QOL) in a Brazilian population. Int Urogynecol J. 2011;22(6):751–5. https://doi.org/10.1007/s00192-010-1354-3.
40. Nojomi M, Digesu GA, Khullar V, Morovatdar N, Haghighi L, Alirezaei M, et al. Validation of Persian version of the prolapse quality-of-life questionnaire (P-QOL). Int Urogynecol J. 2012;23:229–33. https://doi.org/10.1007/s00192-011-1529-6.
41. Morovatdar N, Hghighi L, Najmi Z, Hashemi A, Nojomi M. Response validity of Persian version of P-QOL questionnaire in patients with prolapse. Eur J Obstet Gynecol Reprod Biol. 2015;193:88–91. https://doi.org/10.1016/j.ejogrb.2015.07.013.
42. Chuang FC, Chu LC, Kung FT, Huang KH. Validation of the traditional Chinese version of the prolapse quality of life questionnaire (P-QOL) in a Mandarin-speaking Taiwanese population. Taiwan J Obstet Gynecol. 2016;55(5):680–5. https://doi.org/10.1016/j.tjog.2016.02.018.
43. 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;11:509–21.
44. Cano SJ, Hobart JC. The problem with health measurement. Patient Prefer Adherence. 2011;5:279–90. https://doi.org/10.2147/PPA.S14399.
45. Drost EA (2011) Validity and reliability in social science research. Educ Res Perspect 38 (1):105. Availability: https://search.informit.com.au/documentSummary;dn=491551710186460;res=IELHSS ISSN:0311-2543. [cited 20 Apr 2019].
46. Suhr DD. Exploratory or confirmatory factor analysis? doi: [Internet]. Stat. Data Anal. 2006 [cited 2018 Sep 9]. p. 1–17. 2006. Available from: http://www2.sas.com/proceedings/sugi31/200-31.pdf.
47. Prudon P. Confirmatory factor analysis: a brief introduction and critique. Tilgængelig på. doi:[Internet]. 2011 [cited 2018 Sep 9]. p. 1–18. 2014. Available from: https://pdfssemanticscholar.org/80cc/e29fc22d99c4273bee96ebaa6447c64e268d.pdf.
48. Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. BMC Med Res Methodol. 2011;11(1):87.