Asian Nursing Research 14 (2020) 267–276

Contents lists available at ScienceDirect

Asian Nursing Research

journal homepage: www.asian-nursingresearch.com

Invited Review Article

Evaluation of Studies on the Measurement Properties of Self-Reported Instruments

Eun-Hyun Lee,* Eun Hee Kang, Hyun-Jung Kang

Graduate School of Public Health, Ajou University, Suwon, Republic of Korea

ARTICLE INFO

Article history:
Received 5 November 2020
Received in revised form
25 November 2020
Accepted 25 November 2020

Keywords:
instrument
psychometrics
questionnaire

SUMMARY

Purpose: The purpose of this study was to evaluate studies on the measurement properties of self-reported instruments.
Method: This descriptive review included studies on measurement properties that were reported in *Asian Nursing Research* over a five-year period from 2016 to September 2020. Nine key measurement properties were reviewed for each study: content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, hypotheses-testing construct validity, and responsiveness.
Results: The most commonly applied measurement properties were structural validity and internal consistency. However, structural validity using confirmatory factor analysis or item response theory/Rasch analysis needs to be rigorously analyzed and interpreted. None of the studies assessed measurement error and responsiveness.
Conclusion: It is recommended for nursing researchers to assess measurement properties beyond structural validity and internal consistency using more rigorous methodologies.

Introduction

Clinical nurses and researchers frequently measure objective parameters (e.g., blood pressure) and subjective parameters [e.g., health-related quality of life (HRQOL)] during the work they perform in the field of nursing. Unlike objective parameters such as blood pressure, the subjective parameters cannot be directly measured, and so that they are usually measured using a self-reported instrument comprising questions about the attributes of the subjective parameter being measured. In medicine, a patient-reported outcome (PRO) is a report from patients about a health-related condition and its treatment that is not interpreted by anyone else, such as a clinician [1]. Cappelleri et al. [2] noted that the term of PRO is not limited to patients, being sufficiently broad to also include healthy persons. A tool or instrument for measuring a PRO using a self-reported questionnaire is called a patient-reported outcome measure (PROM).

When using self-reported questionnaires or PROMs, it is very important to consider whether the measurement properties are satisfied. Polit [3] criticized that nurse researchers have mainly focused on the classical measurement ideas of properties that were established decades ago by psychometricians, even though new measurement ideas have evolved and been applied in other health fields. The group for the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) conducted an international Delphi study with 57 experts to reach consensus on the terminology of measurement properties and suggested the following nine key properties for PROMs: content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, hypotheses-testing construct validity, and responsiveness [4,5]. The purpose of the present descriptive study was to determine the extent to which psychometric studies on nursing research have reflected these key measurement properties.

Eun-Hyun Lee: https://orcid.org/0000-0001-7188-3857; Eun Hee Kang: https://orcid.org/0000-0001-9709-7328; Hyun-Jung Kang: https://orcid.org/0000-0001-8374-036X

* Correspondence to: Eun-Hyun Lee, RN, PhD,Graduate School of Public Health, Ajou University, 164 Worldcup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-380, Republic of Korea.

E-mail addresses: ehlee@ajou.ac.kr, hl2dvr@hanmail.net, dusehthsu@gmail.com

https://doi.org/10.1016/j.anr.2020.11.004
p1976-1317 e2093-7482/

Table 1 *Characteristics of the included instruments and studies.*

Concept being measured	Instrument		Study							
	Instrument abbreviation	Target population	Sample size	Age, years Mean ± SD or range	Female, %	Study population	Setting	Country	Language	Original/ translated version
Daytime sleepiness	PDSS-T [6]	Children and adolescents	522	14.0 ± 1.8	50.6	Children	School	Turkey	Turkish	Translated
Quality of life	SQOLPOP [7]	Adolescents diagnosed with cancer†	184	14.6 ± 1.4	47.8	Adolescents diagnosed with cancer	Clinic	Turkey	Turkish	Original
Diabetes: fear of injecting and self-testing	D-FISQ [8]	Patients with diabetes	350	47.3 ± 15.1	54.3	Patients with diabetes	Clinic	Turkey	Turkish	Translated
Self-efficacy of evidence-based practice	K-SE-EBP [9]	Nurses	214	23.5 ± 1.5	92.9	Nurses	Clinic	Korea	Korean	Translated
Wound-specific HRQOL	S-CWIS [10]	Patients with diabetic leg/ foot ulcers	140	58.2 ± 10.0	49.0	Patients with diabetic leg/foot ulcers	Clinic	Sri Lanka	Sinhala	Translated
Sleep disturbance	K-MSQ-Insomnia [11]	College students	470	21.4 ± 2.0	93.0	College students in nursing	School	Korea	Korean	Translated
Health literacy	HL-SF12 [12]	General public	403	44.9 ± 15.8	61.1	Patients	Clinic	Taiwan	Chinese	Translated
Parental–fetal attachment	K-PAFAS [13]	Males with a pregnant spouse	200	unclear	0	Males with a pregnant spouse	Community	Korea	Korean	Original
Stress	T-SNSI [14]	Nursing students	152			Nursing students	School	Turkey	Turkish	Translated
Infertility-related self-efficacy	K-ISE [15]	Infertility patients	314	60.8% were ≥35	74.5	Infertility patients	Clinic	Korea	Korean	Translated
Post-traumatic growth	P-PTGI [16]	Patients with cancer	272	52.7 ± 15.5	59.2	Patients with cancer	Clinic	Iran	Persian	Translated
Decision conflict	K-DCS [17]	Elderly	273	77.3 ± 8.2	80.2	Community-dwelling elderly	Community	Korea	Korean	Translated
Cultural competence	CCSN-SF [18]	Nurses	277 365	29.4 ± 5.7 31.0 ± 7.5	97.1 97.5	Nurses	School Clinic	Korea	Korean	Short form of original version
Bullying	BBNE [19]	Nursing students	442	21.2 ± 1.6	86.0	Nursing students	School	Turkey	Turkish	Partially translated and adopted
Social support	T-CASSS-HB [20]	Children and Adolescents	860	11–14	50.5	Adolescents	School	Turkey	Turkish	Translated
Dietary sodium restriction	DSRQ-I [21]	Patients with hypertension	135	58.2 ± 10.4	54.1	Patients with hypertension	Clinic	Indonesia	Bahasa Indonesia	Translated
Symptoms of negative emotions	K-DASS-21/-12 [22]	Adults	431	39.3 ± 12.1	77.8	Adults	Health center	Korea	Korean	Translated
Research utilization	M-RUQ [23]	Nurses	504/362	42.3 ± 14.1/39.7 ± 10.3	70.6/80.4	Nurses	Clinic	Italy	Italian	Translated
Maternal identity	MRAS-Form B [24]	Primiparous adolescent mothers	397	18.0 (14–20)	100.0	Primiparous adolescent mothers	Community	Thailand	Thai	Translated and modified
Inpatient dignity	IPDS [25]	Patients	Japan, 165; Singapore, 363; UK, 499	>18	33.9, 33.6, 60.3	Patients	Clinic	Japan, Singapore, UK	Japanese, English	Original/ translated
Person-centered perioperative nursing	PCPON [26]	Nurses	459	31.8 ± 7.8	96.9	Nurses	Clinic	Korea	Korean	Original
Drinking behavior	CRAFFT [27]	Adolescents	8,568	15.9 ± 1.5	41.7	Adolescents	School	Korea	Korean	Translated and modified
Physical-activity motivation	C-BREQ-2 [28]	General public	204	79.6 ± 8.7	55.4	Nursing-home residents	Nursing home	China	Chinese	Translated

Health numeracy	DHNT [29]	Patients with diabetes	257	59.8 ± 12.2	45.9	Patients with diabetes	Clinic	Korea	Korean	Original
ICU experience	K-ICEQ [30]	Patients treated in ICUs	200	62.0 ± 12.8	40.5	Patients treated in ICUs	Clinic	Korea	Korean	Translated
Knowledge and attitude regarding pain management	KASP-K [31]	Nurses	81	32.7 ± 9.3	92.6	Nurses in long-term care settings	Clinic	Korea	Korean	Translated and modified
Work control	C-WCS [32]	Workers	718	31.1 ± 6.9	98.1	Nurses	Clinic	China	Chinese	Translated and modified

PDSS-T, Turkish version of Pediatric Daytime Sleepiness Scale; SQOLPOP, Scale for Quality of Life in Pediatric Oncology Patients; D-FISQ, Diabetes Fear of Injecting and Self-testing Questionnaire; K-SE-EBP, Korean version of Self-Efficacy of Evidence-Based Practice; S-CWIS, Sinhala version of Cardiff Wound Impact Schedule; K-MSQ-Insomnia, Korean version of Mini-Sleep Questionnaire-Insomnia; HL-SF12, 12-item Short-Form Health Literacy questionnaire derived from HLS-EU-Q (European Health Literacy Survey Questionnaire); K-PAFAS, Korean Parental–Fetal Attachment Scale; T-SNSI, Turkish version of Student Nurse Stress Index; K-ISE, Korean version of Infertility Self-Efficacy; P-PTGI, Persian version of Posttraumatic Growth Inventory; K-DCS, Korean version of Decisional Conflict Scale; CCSN-SF, Short Form of the Cultural Competence Scale for Nurses; BBNE, Bullying Behaviors in Nursing Education adopted from the WPVB (Workplace Psychological Violence Behaviors); T-CASSS-HB, Turkish version of Child and Adolescent Social Support Scale for Healthy Behaviors; DSRQ-I, Indonesian version of Dietary Sodium Restriction Questionnaire; K-DASS-21/-12, Korean versions of Depression Anxiety Stress Scales-21 and -12; M-RUQ, Modified Research Utilization Questionnaire; MRAS-Form B, Maternal Role Attainment Scale Form B; IPDS, Inpatient Dignity Scale; PCPON, Person-Centered Perioperative Nursing Scale; CRAFFT, Car, Relax, Alone, Forget, Family/Friends, Trouble; C-BREQ-2, Chinese version of Behavioral Regulation in Exercise Questionnaire-2; DHNT, Diabetes Health Numeracy Test; K-ICEQ, Korean version of Intensive-Care Experience Questionnaire; KASP-K, Knowledge and Attitudes Survey on pain management for Korean long-term care professionals; C-WCS, Chinese version of Work Control Scale; HRQOL, health-related quality of life; ICU, intensive-care unit.
SD, standard deviation.
† The parental form of the instrument was excluded because it was a proxy measurement.

Method

Studies on the measurement properties of self-reported instruments were selected from *Asian Nursing Research* that had been published over a five-year period, which comprised a sample of 186 studies published from 2016 to September 2020. These studies included 29 psychometric studies, of which two were eliminated due to proxy measurements, and therefore 27 studies were reviewed based on the nine measurement properties suggested by the COSMIN.

Results and discussion

General characteristics of the included instruments and studies

Table 1 presents the 27 studies in which the identified instruments were evaluated [6–32]. The studies were conducted in clinic, community, and/or school settings, and they had sample sizes ranging from 81 to 8,568 participants. The largest proportion of studies were conducted in South Korea ($n = 12$, 44.4%), followed by Turkey ($n = 6$, 22.2%) and China ($n = 3$, 11.1%), with one study in each of Indonesia, Italy, Thailand, Iran, and Sri Lanka. In addition, there was one international collaboration study involving Japan, Singapore, and the UK. Five studies (18.5%) developed a new instrument [7,13,25,26,28], while the remaining tested translated or short versions of existing instruments.

Content validity

Content validity is the most important measurement property of an instrument, because it may affect the likelihood of the instrument fulfilling other measurement properties [33]. Content validity is defined as the degree to which the content of an instrument adequately reflects the construct being measured [4], in terms of the three aspects of relevance, comprehensiveness, and comprehensibility [34]. In the present study, 21 of the included studies (77.8%) assessed the relevance using a content validity index, although the method used for the relevance was not clearly reported for 2 of them [24,25]. Comprehensibility and comprehensiveness were assessed in 48.2% ($n = 14$) and 3.7% ($n = 1$) of the studies, respectively (Table 2).

The availability of a definition of the construct being measured should be a prerequisite for the content validity of a new instrument. In this study, conceptual definitions were clearly reported for only two [26,29] of the five newly developed instruments (Tables 1 and 2). The systematic review of a disability index by Ailliet et al. [35] revealed that the instrument was unclear regarding what it aims to measure and recommended developing a new neck-specific instrument with a clear definition.

Internal structure

The structural validity, consistency, and cross-cultural validity/measurement invariance were measured to determine the internal structure of each self-reported instrument. Assessments of structural validity should be preceded by assessments of the internal consistency or cross-cultural validity/measurement invariance [5].

Structural validity

Structural validity is defined as the degree to which the scores of a measurement instrument adequately reflect the dimensionality of the construct being measured [4], and it can be assessed using factor analysis and item response theory (IRT)/Rasch analysis. The present study (Table 2) found that factor analysis was performed in

Table 2 *Content validity, structural validity, internal consistency, and cross-cultural validity/measurement invariance.*

Instrument abbreviation	Content validity				Structural validity						Internal consistency	Cross-cultural validity/ measurement invariance
					EFA		CFA or IRT/Rasch analysis					
	D	R	Cb	Cv	Number of factors	Explained variance, %	Fit indices	AVE	Discriminant validity using AVE	Cross-validation with same/ different samples	Composite reliability	
PDSS-T [6]		○			1	41.1	RMSEA = .07, GFI = .97. CFI = .97, NFI = .96, NNFI = .95, IFI = .97			Same		Cronbach α
SQOLPOP [7]		○			1	80.4	χ^2/df = 4.50, GFI = .90, CFI = .91, NFI = .90, IFI = .91, RMSEA = .079			Same		Cronbach α
D-FISQ [8]		○	○		2		χ^2/df = 1.00, GFI = .70, AGFI = .59, CFI = .93, RMSEA = .018					Cronbach α for each subscale
K-SE-EBP [9]					3		CFI = .91, TLI = .90, RMSEA = .075					Cronbach α for each subscale
S-CWIS [10]		○	○		3							Cronbach α for each subscale
K-MSQ-Insomnia [11]					1	56.0						Cronbach α
HL-SF12 [12]					3		χ^2/df = 3.27, RMSEA = .07, GFI = .94, AGFI = .91, CFI = .94, IFI = .94, NFI = .92					Cronbach α for total scale
K-PAFAS [13]		○	○		4	59.0	GFI = .836, AGFI = .790, NFI = .793, CFI = .868, RMSEA = .082			Same		Cronbach α for each subscale
T-SNSI [14]		○			4	67.0	χ^2/df = 1.76, GFI = .89, IFI = .94, RMSE≤ .07, TLI = .92 CFI = .94, SRMR .08			Same		Cronbach α for each subscale
K-ISE [15]		○	○		1	58.4	χ^2/df = 1.08, CFI = .99, NFI = .96, RMSEA = .02, GFI = .94, SRMR = .03			Different		Cronbach α
P-PTGI [16]		○			5		RMSEA = .10, NFI = .93,	○	○			Cronbach α for each subscale

					NNFI = .94, CFI = .95, IFI = .95, SRMR = .08					
K-DCS [17]			2	69.8					Cronbach α for the K-DCS were not reported	
CCSN-SF [18]	○	○	4	63.3	GFI = .98, AGFI = .97, NFI = .97, SRMR = .06	○	○	Different	Cronbach α for each subscale	
BBNE [19]		○	4		χ^2/df = 2.60, RMSEA = /06, SRMR = .05, RMR = .06, NFI = .95, NNFI = .96, CFI = .97, GFI = .93, AGFI = .90	○		○	Cronbach α for each subscale	
T-CASSS-HB [20]	○		5	Frequency, 76.0; importance, 90.0	Frequency/ importance: χ^2/df = 4.67/ 1.95, CFI = .97/ .99, SRMR = .02/.02, RMSEA = .06/ .03			Same	Cronbach α for each subscale	
DSRQ-I [21]	○	○	3	64.2	χ^2/df = 2.17, GFI = .85, AGFI = .79, SRMR = .07, RMSEA = .09, CFI = .90, NFI = .83			Same	Cronbach α for each subscale	
K-DASS-21/-12 [22]	○	○	3		21/12 versions: SRMR = .049/ .034, GFI = .883/.934, CFI = .903/.960 RMSEA = .072/ .079				Cronbach α for each subscale	
M-RUQ [23]	○	○	3	49.6	CFI = .91, RMSEA = .051, SRMR = 1.00			Different	Cronbach α for each subscale	
MRAS-Form B [24]	○		3		χ^2/df = 2.23, CFI = .93, TLI = .92, RMSEA = .06, SRMR = .05	○		○	Cronbach α for each subscale	
IPDS [25]	○	○	4	Expectation, 54.5; satisfaction, 58.5	Singaporean population, Expectation: χ^2/df = 2.85 SRMR = .05 CFI = .94 RMSEA = .08 Satisfaction: χ^2/df = 2.23 SRMR = .03	○		Same	Cronbach α for each subscale	UK population, Expectation: χ^2/df = 2.32 SRMR = .06 CFI = .92 RMSEA = .09 Satisfaction: χ^2/df = 2.60 SRMR = .07 CFI = .86 RMSEA = .10

(*continued on next page*)

Table 2 (*continued*)

Instrument abbreviation	Content validity				Structural validity						Internal consistency	Cross-cultural validity/ measurement invariance
					EFA		CFA or IRT/Rasch analysis					
	D	R	Cb	Cv	Number of factors	Explained variance, %	Fit indices	AVE	Discriminant validity using AVE	Cross-validation with same/ different samples	Composite reliability	
PCPON [26]	○	○	○		5	68.2	CFI = .96 RMSEA = .06 χ^2/df = 1.65, GFI = .90, RMR = .03, SRMR = .06, RMSEA = .05, TLI = .92, CFI = .93	○	○	Different	○	Cronbach α for each subscale	
CRAFFT [27]					1		IRT with two-parameter logistic model. No information on assumption values for unidimensionality, local independence, or monotonicity tests. χ^2 fit statistic was given					KR-20	Significant DIF in 4 items out of 6 items by gender
C-BREQ-2 [28]		○			5		CFI = .94, SRMR = .05, RMSEA = .07	○				Cronbach α for each subscale	
DHNT [29]	○	○	○	○	1	56.6	Rasch analysis: unidimensionality was reported. Model fit was satisfied					KR-20	
K-ICEQ [30]		○	○		4		χ^2/df = 1.87, TLI = .90, CFI = .91, RMSEA = .06	○			○	Cronbach α for each subscale	
KASP-K [31]		○			Unknown		Stated that IRT was used, but appropriate model fit values were not reported					Cronbach α for all items rather than KR-20	
C-WCS [32]		○			3	55.1	χ^2/df = 2.63, TLI = .89, CFI = .91, SRMR = .06, RMSEA = .07	○	○	Different	○	Cronbach α for each subscale	

D, definition; R, relevance; Ch, comprehensibility; Cv, comprehensiveness; EFA, exploratory factor analysis; CFA, confirmatory factor analysis; AVE, average variance extracted; IRT, item response theory; RMSEA, root-mean-square error of approximation; TLI, Tucker-Lewis index; SRMR, standardized root-mean-square residual; GFI, goodness-of-fit index; AGFI, adjusted goodness-of-fit index; CFI, comparative fit index; IFI, incremental fit index; NFI, normative fit index; NNFI, non-normed fit index; DIF, differential item functioning; KR-20, Kuder–Richardson–20.
○, measurement property was assessed.

23 studies (85.2%), IRT/Rasch analysis was used in three studies (11.1%), and one study did not assess structural validity.

There are two types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is normally used to reduce the number of items or to explore the number of factors of a new instrument in the absence of a prior hypothesis. CFA is used to test a hypothesized factorial structure based on a theory or previous empirical evidence [36]. This means that CFA is more appropriate than EFA for assessing the structural validity of an instrument if there is already information available on the dimensionality of the instrument. If there is no (or little) information available on the structure of a construct to be assessed, it is recommended to conduct EFA to identify the structure, and then CFA can be used to confirm whether or not the structure provides a good fit.

The present study found that factor analyses for structural validity were undertaken in 23 studies (92.0%) of the 25 included studies. Two studies (8.7%) undertook EFA alone, nine studies (39.1%) used CFA alone, and 12 studies (52.2%) utilized both EFA and CFA. Five [15,18,23,26,32] of the 12 studies that used both EFA and CFA used a cross-validation approach with different (sub)samples for EFA and CFA to prevent the reflection of idiosyncrasies (Table 2). The proportions of the factor analyses that used CFA alone or both CFA and EFA were higher than for those conducted in the psychometric studies published in three nursing journals from 2010 to 2014 (Factor analyses were undertaken in 81.0%, $n = 85$). EFA was performed alone in 60.0% of the factor analyses, CFA was performed alone in 21.2%, and both EFA and CFA were performed in 18.8% [3]. The higher percentages might be due to the studies included in the present analysis mainly assessing translated versions of existing instruments (81.5%). In other words, the researchers who performed the studies included here could have had access to information on the structural instruments from the original versions, thereby avoiding the need to apply EFA.

Even though most of the studies analyzed here applied factor analyses, the details need to be carefully checked. Regarding EFA, the total variance was not explained by all items based on applying a criterion of >50% [37] in two [6,23] of the 15 studies. For one study [31], the authors noted that factor analysis could not be conducted because the item response was a binary type (correct/incorrect). As the authors asserted, utilizing factor analysis in such a situation is problematic because it uses a product-moment correlation matrix based on the assumption that the responses conform to a normal distribution. However, EFA is possible if a tetrachoric correlation matrix is used, which is applicable to binary response items [36]. For example, Lee et al. [29] performed EFA using SPSS syntax software with a tetrachoric correlation matrix to identify the dimensionality of the Diabetes Health Numeracy Test. Another approach is to use other programs that are suitable for binary response items (e.g., Mplus).

Regarding CFA using structural equation modeling (SEM), the goodness of fit of a measurement model can be assessed using numerous indexes, such as absolute fit indexes (χ^2/df, root-mean-square error of approximation, standardized root-mean-square residual, goodness-of-fit index, and adjusted goodness-of-fit index) and comparative fit indexes (comparative fit index, incremental fit index, normative fit index, and non-normed fit index). One of the included studies [26] provided information about the acceptable criterion values for the fit indexes but did not include supporting citations for the values used, whereas another study [13] gave supporting citations for the criterion values, but most of the fit index values were not satisfied or interpreted using an acceptable measurement model. A measurement model may be modified to improve its model fit, such as by applying modified indexes [38]. A model modification allowing error covariance between two items was applied in seven studies [9,14,19,22,24,30,32]. Only one study [22] reported a substantial improvement in model fit using a significant difference in the χ^2 values between the original model and its modified model.

As an ancillary analysis, the average variance extracted (AVE) is often reported in CFA using SEM, which refers to the average of the squared standardized pattern coefficients for the items within a subscale [39]. Even though popular SEM programs (e.g., AMOS) do not provide the AVE value, this is easy to calculate manually. For nine of the included studies, the AVE values was reported as the convergent validity of items under a construct (subscale) with a criterion of >.50 [40]. However, Kline [39] noted that the AVE is based on a standardized coefficient, and therefore it might not be suitable for comparing the same items across different samples. The composite reliability based on unstandardized coefficients is preferred because it can be compared across samples.

The AVE can be applied to demonstrate discriminant validity between constructs (subscales) compared with the shared variance between the subscales [40]. If the squared coefficient of the correlation between subscales is greater than the AVE values for at least one of the subscales, the two subscales are not distinguishable. This was reported for six of the included studies. In two studies [24,30], at least one AVE of the two subscales was less than its highest squared coefficient of the correlations with the subscales; in other words, the subscales were not distinguishable, but the nondistinguishability problem was left without any solution. In addition to those two studies, another study [12] conducted CFA and found satisfactory goodness-of-fit indices, with 12 items loading onto three factors of an instrument measuring health literacy. However, the coefficients for the correlations between pairs of factors were high ($\varphi = .80$, .92 and .94), reflecting the incorporation of the factors [41]. Based on the coefficients for the correlations among factors reported on in the study, shared variances (φ^2) were calculated as .64, .85, and .88, respectively. All of these values were greater than the AVE values of the three factors (which were not directly reported by the authors, but they are easy to calculate manually), requiring follow-up handling.

The distinguishability problem between factors in CFA can be handled in a few ways: (1) inquiring about cross-loading items and determining which items to eliminate, (2) combining the factors and then comparing the combined measurement model with the original model, or (3) exploring a higher-order model or a bifactor model. For example, Lee and Lee [42] assessed the structural factor for the first-order three-factor model of the original scale to apply in another population. The CFA performed in their study revealed that two factors were poorly distinguishable, and therefore these two factors were combined into a single factor. However, the two-factor model did not represent a significant improvement over the three-factor model. As the next solution, a second-order three-factor model was assessed and compared with the first-order three-factor model of the scale.

Regarding the AVE and discriminant validity between subscales, three studies [18,30,32] addressed this in the same way as hypothesis-testing validity, such as by determining the convergent and discriminant validity of an instrument (which is explained below). It needs to be precisely described that all of these methods are used to assess the structural validity of a measurement model in CFA.

IRT/Rasch analysis provides rich information about individual items that is not available using classical test theory, and therefore it has been recently applied for the assessment of PROMs. Three studies applied IRT/Rasch analysis to structural validity. However, for one study [27] there were no values reported for the assumption tests for IRT. For another of the studies [29], model fit values (infit and outfit mean squares) were reported for the assumption test of

Table 3 *Remaining measurement properties.*

Instrument abbreviation	Reliability			Criterion validity		Hypotheses-testing construct validity		
	Interval	Correlation	ICC	Concurrent	Predictive	Convergent validity	Discriminant validity	Known-groups validity
PDSS-T [6]								○
SQOLPOP [7]	3 weeks	○						
D-FISQ [8]	2 weeks		○					○
K-SE-EBP [9]								
S-CWIS [10]	2 weeks		○	○				○
K-MSQ-Insomnia [11]	1 week		○	○	○			
HL-SF12 [12]						○		○
K-PAFAS [13]	2 weeks	○		○		○	○	
T-SNSI [14]	2 weeks	○						
K-ISE [15]	2-3 weeks		○	○				
P-PTGI [16]	4 weeks	○						
K-DCS [17]	2 weeks	○		○				
CCSN-SF [18]						○		○
BBNE [19]								
T-CASSS-HB [20]	6 weeks		○					
DSRQ-I [21]								
K-DASS-21/-12 [22]				○		○	○	○
M-RUQ [23]	20 days		○					
MRAS-Form B [24]								
IPDS [25]				○				
PCPON [26]				○				
CRAFFT [27]								
C-BREQ-2 [28]	7 days		○					
DHNT [29]				○		○		
K-ICEQ [30]								○
KASP-K [31]	2 weeks	○						○
C-WCS [32]	2 weeks		○					

ICC, intraclass correlation coefficient; ○, measurement property was assessed.

unidimensionality, while no IRT-related values were reported for the third study [31].

Internal consistency

All except one [17] of the studies (Table 2) assessed internal consistency and determined Cronbach's α or Kuder–Richardson–20 (KR-20) values. One study [12] provided only Cronbach's α for the total items, despite the instrument comprising three subscales. If the multiple subscales of an instrument were identified by structural validity (except for a higher-order or bifactor CFA), Cronbach's α for the total items might be ignored because the internal consistency of each subscale is relevant [5]. The KR-20 value is appropriate only for an instrument with binary responses (e.g., correct/incorrect or yes/no) [33]. One study used Cronbach's α even though the items had binary responses [31].

In the study involving the Korean version of the Decisional Conflict Scale (K-DCS) [17], Cronbach's α was assessed using the original 16-item DCS comprising five subscales, and three subscales satisfied internal consistency. EFA was performed using only the nine items that satisfied internal consistency of the three subscales, and the K-DCS was extracted comprising two subscales (and this was not followed by determining Cronbach's α for the newly extracted two subscales in a Korean population). In the tests, the authors eliminated all items of the two subscales that did not satisfy a Cronbach's α value of .70. However, it should be remembered that good internal consistency does not guarantee the presence of good structural validity.

Cross-cultural validity/measurement invariance

Cross-cultural validity refers to the degree to which the performance of the items in a translated or culturally adapted instrument adequately reflects the performance of the items in the original version of the instrument [5]. Therefore, at least two groups are required (e.g., language, country, gender, and age groups). This validity is usually assessed using multiple-group CFA or differential item functioning (DIF). Two of the analyzed studies (Table 2) assessed cross-cultural validity/measurement invariance. One study [25] developed a scale to measure the dignity of in-patients, which is applicable in a cross-cultural context. Therefore, data were collected in Singaporean and UK populations, and CFA was applied separately to the data from each population. In this case, the application of multiple-group CFA is recommended for investigating structural invariance between the cultural groups rather than separately conducting CFA in each population. Another study [27] used DIF to investigate item invariance by gender based on IRT.

Reliability

Reliability was assessed in 14 of the studies, by administering the same instruments to the same respondents at different times (i.e., the test–retest reliability) (Table 3). The intraclass correlation coefficient (ICC) is the preferred statistic when the score is a continuous measure [36]. Eight of the 14 studies (64.3%) used the ICC, while the remaining studies used Pearson's correlation coefficient, which informs about the association between two variables. One study [7] used the coefficient of the correlation between the items using a five-point Likert scale for the test–retest reliability.

When assessing the test–retest reliability, the time interval between the repeated measurements needs to be sufficiently long to prevent recall bias but not too long to allow changes to occur in the characteristics of the respondents related to the construct being measured [33]. Without a specific reason, an interval of about 2 weeks is generally accepted. All of the studies analyzed in the present study used intervals from 1 to 3 weeks, with the exception of one study that used a six-week interval [20]. When measuring the test–retest reliability, the attributes of the construct being measured should be temporally stable (e.g., cognitive and trait scales). It is not appropriate for a state construct to be expected to

change over time, such as mood [43]. Lee et al. [22] were the only authors who reported that they did not conduct test–retest reliability because the instrument being assessed was a state construct.

Criterion validity

Criterion validity is defined as the degree to which the scores of a measurement instrument adequately reflect a standard [4], and there are two types: concurrent validity (an instrument being validated and a selected criterion measured at the same time) and predictive validity (a criterion is measured in the future). If a gold standard for the construct being measured is not available, criterion validity does not need to be assessed. Mokkink et al. [5] noted that there is no gold standard available for a PROM, although a long version of a PROM can be used as a surrogate for its corresponding short version. From a broader perspective, Polit and Yang [33] noted that a gold standard might be available for some self-reported instruments. For example, a general health-related QOL instrument can be considered as a criterion for a disease-specific QOL instrument [44,45].

In the present study, concurrent validity was assessed in nine studies using continuous measures. Four studies [10,15,22,29] used generic or specific types of criterion for the PROMs being measured, while the remaining studies utilized instruments measuring relevant constructs as their criteria, which might be more appropriate for the assessment of convergent validity. The authors of several studies criticized existing instruments and asserted the need to develop a new instrument in the Introduction section of the corresponding report; however, they then paradoxically used the criticized instruments as the gold standards for assessing criterion validity for their new instruments.

For the predictive validity, an instrument being assessed is administered first, and then a criterion instrument is administered after an appropriate interval. One of the studies analyzed here assessed predictive validity by administering the assessed and criterion instruments simultaneously [11], which corresponds to concurrent validity rather than predictive validity.

Hypotheses-testing construct validity

Hypotheses-testing validity is defined as the relationships of scores on the instrument of interest with the scores on other instruments measuring similar constructs (convergent validity) or dissimilar constructs (discriminant validity), or the difference in the instrument scores between subgroups of people (known-groups validity) [36]. When designing a psychometric study, it is recommend to formulate hypotheses about the expected direction and magnitude of the correlations or differences for the validation. Then, the validation can be performed by analyzing the data regarding whether or not the formulated hypotheses were satisfied. However, many researchers determine the validity only based on the statistical significance of the employed statistics, without considering the expected direction and magnitude.

The present study ignored the tests for convergent and discriminant (known-groups) validity, which was simply performed with the general characteristics as their comparators. Convergent validity was assessed in six studies, with hypotheses being formulated *a priori* in four of them [11,12,22,29]. Two studies examined discriminant validity, but the hypothesis was reported for only one of them [22]. Known-groups validity was assessed in seven studies, with hypotheses being set in four of them [8,10,22,30].

Measurement error and responsiveness

A measurement error is a systematic and random error in a subject's score that is not attributable to true changes in the construct being measured [4]. The preferred statistic for measurement error in continuous scores is the standard error of the measurement based on a test–retest design, smallest detectable change, or the limits of agreement [34,46]. None of the studies analyzed in the present study assessed measurement errors.

The measurement property of responsiveness refers to the ability of a PROM to detect changes over time in the construct being measured [4]. This requires a longitudinal research design in which participants have to respond at least twice on the instrument so that it is validated over at least one interval. An event or condition (e.g., treatment or intervention) that is known to induce a score change of the instrument construct must be applied during the interval. For example, for the responsiveness of the asthma-specific QOL, Lee et al. [47] administered the instrument to newly diagnosed patients twice at baseline and 1 month later after they had been treated. Responsiveness tends to be rarely assessed with such a longitudinal design, and none of the studies analyzed in the present study assessed responsiveness.

Conclusions

This paper has reviewed studies on the measurement properties of self-reported instruments published in *Asian Nursing Research* over the last 5 years. The most frequently tested measurement properties were structural validity and internal consistency. However, the findings for structural validity assessed using CFA or IRT/Rasch analysis need to be rigorously analyzed and interpreted. Most assessments of the content validity focused on the item relevance of the construct being measured, with comprehensibility and comprehensiveness rarely being covered. For criterion validity, the selection of a gold standard should be carefully considered. For the hypotheses-testing construct validity, it is recommended to formulate the expected relationships or differences with other comparator instruments or groups when designing a study. In particular, a longitudinal design needs to be considered for assessing the test–retest reliability or responsiveness.

Together the present findings recommend further assessments of measurement properties beyond structural validity and internal consistency using more rigorous methodologies. It should be noted that the findings of this study were based on a sample of studies from a single journal, and thus they cannot be generalized to studies reported on in other nursing journals.

Author contributions

E-H.L. conceived the study and wrote the manuscript. E-H.L., E.H.K., and H.K. analyzed the included studies and contributed to and approved the final submitted version of the manuscript.

Ethical approval

Not required.

Funding

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2018R1A2B6001719). The funder did not play any role in the conduct or publication of the study.

Conflict of interest

One author (E.H.L.) was involved in two of the studies analyzed in the present study. The other authors have no conflict of interest to declare.

References

1. Weldring T, Smith SMS. Article commentary: patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights. 2013;6. https://doi.org/10.4137/HSI.S11093. HSI.S11093.
2. Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T. Patient-reported outcomes: measurement, implementation, and interpretation. Florida: CRC Press Taylor & Francis Group; 2014.
3. Polit DF. Assessing measurement in health: beyond reliability and validity. Int J Nurs Stud. 2015;52(11):1746–53. https://doi.org/10.1016/j.ijnurstu.2015.07.002
4. 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. https://doi.org/10.1016/j.jclinepi.2010.02.006
5. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1171–9. https://doi.org/10.1007/s11136-017-1765-4
6. Bektas M, Bektas I, Ayar D, Selekoglu Y, Ayar U, Kudubes AA, et al. Psychometric properties of Turkish version of pediatric daytime sleepiness scale (PDSS-T). Asian Nurs Res. 2016;10(1):62–7. https://doi.org/10.1016/j.anr.2016.01.002
7. Bektas M, Akdeniz Kudubes A, Ugur O, Vergin C, Demirag B. Developing the scale for quality of life in pediatric oncology patients aged 13–18: adolescent form and parent form. Asian Nurs Res. 2016;10(2):106–15. https://doi.org/10.1016/j.anr.2016.03.002
8. Celik S, Pinar R. Psychometric evaluation of a Turkish version of the diabetes fear of self-injecting and self-testing questionnaire (D-FISQ). Asian Nurs Res. 2016;10(3):195–200. https://doi.org/10.1016/j.anr.2016.06.001
9. Oh EG, Yang YL, Sung JH, Park CG, Chang AM. Psychometric properties of Korean version of self-efficacy of evidence-based practice scale. Asian Nurs Res. 2016;10(3):207–12. https://doi.org/10.1016/j.anr.2016.05.003
10. Sriyani KA, Gunawardena N, Wasalathanthri S, Hettiarachchi P. Validation of Sinhala version of Cardiff wound impact schedule in patients with diabetic leg and foot ulcers. Asian Nurs Res. 2016;10(3):240–5. https://doi.org/10.1016/j.anr.2016.06.005
11. Kim H-J. Validation of the Korean version of the mini-sleep questionnaire–Insomnia in Korean college students. Asian Nurs Res. 2017;11(1):1–5. https://doi.org/10.1016/j.anr.2017.01.001
12. Duong TV, Chang PW, Yang S-H, Chen M-C, Chao W-T, Chen T, et al. A new comprehensive short-form health literacy survey tool for patients in general. Asian Nurs Res. 2017;11(1):30–5. https://doi.org/10.1016/j.anr.2017.02.001
13. Noh NI, Yeom H-A. Development of the Korean paternal-fetal attachment scale (K-PAFAS). Asian Nurs Res. 2017;11(2):98–106. https://doi.org/10.1016/j.anr.2017.05.001
14. Sarikoc G, Bayram Demiralp M, Oksuz E, Pazar B. Turkish version of the student nurse stress index: validity and reliability. Asian Nurs Res. 2017;11(2):128–33. https://doi.org/10.1016/j.anr.2017.05.006
15. Kim JH, Park HJ, Kim JH, Chung S, Hong HJ. Psychometric properties of the Korean version of the infertility self-efficacy scale. Asian Nurs Res. 2017;11(3):159–65. https://doi.org/10.1016/j.anr.2017.06.002
16. Heidarzadeh M, Naseri P, Shamshiri M, Dadkhah B, Rassouli M, Gholchin M. Evaluating the factor structure of the Persian version of posttraumatic growth inventory in cancer patients. Asian Nurs Res. 2017;11(3):180–6. https://doi.org/10.1016/j.anr.2017.07.003
17. Kim J, Kim S, Hong SW, Kang S-W, An M. Validation of the decisional conflict scale for evaluating advance care decision conflict in community-dwelling older adults. Asian Nurs Res. 2017;11(4):297–303. https://doi.org/10.1016/j.anr.2017.11.004
18. Chae D, Park Y. Development and cross-validation of the short form of the cultural competence scale for nurses. Asian Nurs Res. 2018;12(1):69–76. https://doi.org/10.1016/j.anr.2018.02.004
19. Cerit K, Türkmen Keskin S, Ekici D. Development of instrument of bullying behaviors in nursing education based on structured equation modeling. Asian Nurs Res. 2018;12(4):245–50. https://doi.org/10.1016/j.anr.2018.07.002
20. Albayrak S, Çakır B, Kılınç FN, Vergili Ö, Erdem Y. Reliability and validity study of the Turkish version of child and adolescent social support scale for healthy behaviors. Asian Nurs Res. 2018;12(4):273–8. https://doi.org/10.1016/j.anr.2018.10.004
21. Wicaksana AL, Wang S-T. Psychometric testing of the Indonesian version of dietary sodium restriction questionnaire among patients with hypertension. Asian Nurs Res. 2018;12(4):279–85. https://doi.org/10.1016/j.anr.2018.10.005
22. Lee E-H, Moon SH, Cho MS, Park ES, Kim SY, Han JS, et al. The 21-item and 12-item versions of the depression anxiety stress scales: psychometric evaluation in a Korean population. Asian Nurs Res. 2019;13(1):30–7. https://doi.org/10.1016/j.anr.2018.11.006
23. Caruso R, Grugnetti AM, Pastore U, Dellafiore F, Pittella F, Ausili D, et al. Modified research utilization questionnaire: development and validation study among Italian nurses. Asian Nurs Res. 2019;13(1):61–8. https://doi.org/10.1016/j.anr.2019.01.006
24. Panthumas S, Kittipichai W. Validation of the maternal identity scale for primiparous Thai teenage mothers. Asian Nurs Res. 2019;13(1):69–75. https://doi.org/10.1016/j.anr.2019.01.007
25. Ota K, Maeda J, Gallagher A, Yahiro M, Niimi Y, Chan MF, et al. Development of the inpatient dignity scale through studies in Japan, Singapore, and the United Kingdom. Asian Nurs Res. 2019;13(1):76–85. https://doi.org/10.1016/j.anr.2019.01.008
26. Shin S, Kang J. Development and validation of a person-centered perioperative nursing scale. Asian Nurs Res. 2019;13(3):221–7. https://doi.org/10.1016/j.anr.2019.07.002
27. Song Y, Kim H, Park S-Y. An item response theory analysis of the Korean version of the CRAFFT scale for alcohol use among adolescents in Korea. Asian Nurs Res. 2019;13(4):249–56. https://doi.org/10.1016/j.anr.2019.09.003
28. Liu L, Xiang M, Guo H, Sun Z, Wu T, Liu H. Reliability and validity of the behavioral regulation in exercise questionnaire-2 for nursing home residents in China. Asian Nurs Res. 2020;14(1):11–6. https://doi.org/10.1016/j.anr.2019.12.002
29. Lee E-H, Lee YW, Lee K-W, Hong S, Kim SH. A new objective health numeracy test for patients with type 2 diabetes: development and evaluation of psychometric properties. Asian Nurs Res. 2020;14(2):66–72. https://doi.org/10.1016/j.anr.2020.01.006
30. Kang J, Cho YS. Cross-cultural validation of the intensive care experience questionnaire in Korean critical care survivors. Asian Nurs Res. 2020;14(2):89–96. https://doi.org/10.1016/j.anr.2020.03.002
31. Kwon S-H, Kim H, Park S, Jeon W. Development of knowledge and attitudes survey on pain management for Korean long-term care professionals. Asian Nurs Res. 2020;14(2):105–13. https://doi.org/10.1016/j.anr.2020.04.002
32. Wen F, Wang L, Fang Z, Zhu J, Zhang Y. The Chinese version of the work control scale for nurses: modifying the translation and psychometric testing. Asian Nurs Res. 2020;14(2):122–8. https://doi.org/10.1016/j.anr.2020.04.006
33. Polit DF, Yang FM. Measurement and the measurement of change. Philadelphia: Wolters Kluwer; 2016.
34. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159–70. https://doi.org/10.1007/s11136-018-1829-0
35. Ailliet L, Knol DL, Rubinstein SM, de Vet HCW, van Tulder MW, Terwee CB. Definition of the construct to be measured is a prerequisite for the assessment of validity. The neck disability index as an example. J Clin Epidemiol. 2013;66(7). https://doi.org/10.1016/j.jclinepi.2013.02.005, 775–782.e2.
36. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. London: Cambridge University Press; 2011. p. 77–80, 175–181.
37. Pett MA, Lackey NR, Sullivan JJ. Making sense of factor analysis. California: Sage; 2003. p. 90–101.
38. Byrne BM. Structural equation modeling with AMOS: basic concepts, applications, and programming. New York: Routledge; 2016. p. 86–9.
39. Kline RB. Principles and practice of structural equation modeling. 4th ed. New York: The Guilford; 2016.
40. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Market Res. 1981;18(1):39–50. https://doi.org/10.1177/002224378101800104
41. Tabachnick BG, Fidell LS. Using multivariate statistics. Massachusetts: Alleyn & Bacon; 2007.
42. Lee E-H, Lee YW. First-order vs. second-order structural validity of the health literacy scale in patients with diabetes. Scand J Caring Sci. 2017;32(1):441–7. https://doi.org/10.1111/scs.12460
43. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. J Nurs Scholarsh. 2007;39(2):155–64. https://doi.org/10.1111/j.1547-5069.2007.00161.x
44. Talley NJ, Haque M, Wyeth JW, Stace NH, Tytgat GN, Stanghellini V, et al. Development of a new dyspepsia impact scale: the Nepean dyspepsia index. Aliment Pharmacol Ther. 1999;13(2):225–35. https://doi.org/10.1046/j.1365-2036.1999.00445.x
45. Lee E-H, Kwon O, Hahm KB, Kim W, Kim JI, Cheung DY, et al. Irritable bowel syndrome-specific health-related quality of life instrument: development and psychometric evaluation. Health Qual Life Outcome. 2016;14(1):22. https://doi.org/10.1186/s12955-016-0423-9
46. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147–57. https://doi.org/10.1007/s11136-018-1798-3
47. Lee E-H, Kim S-H, Choi J-H, Jee Y-K, Nahm D-H, Park H-S. Development and evaluation of an asthma-specific quality of life (A-QOL) questionnaire. J Asthma. 2009;46(7):716–21. https://doi.org/10.1080/02770900903067887