International Journal of Nursing Studies 79 (2018) 58–69

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/locate/ijns

Evaluating test-retest reliability in patient-reported outcome measures for older people: A systematic review

Myung Sook Park[a], Kyung Ja Kang[b], Sun Joo Jang[c], Joo Yun Lee[d], Sun Ju Chang[e,*]

[a] Nursing Department, Konkuk University, Chungju, South Korea
[b] College of Nursing, Jeju National University, Jeju, South Korea
[c] College of Nursing, Eulji University, Daejeon, South Korea
[d] The Research Institute of Nursing Science, Seoul National University, Seoul, South Korea
[e] College of Nursing & The Research Institute of Nursing Science, Seoul National University, Seoul, South Korea

ARTICLE INFO

Keywords:
Test-retest reliability
Patient-reported outcomes
Systematic review
Aged

ABSTRACT

Objectives: This study aimed to evaluate the components of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported outcome measures in older people and to provide suggestions on the methodology for calculating test-retest reliability for patient-reported outcomes in older people.
Design: This was a systematic literature review.
Data sources: MEDLINE, Embase, CINAHL, and PsycINFO were searched from January 1, 2000 to August 10, 2017 by an information specialist.
Review methods: This systematic review was guided by both the Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist and the guideline for systematic review published by the National Evidence-based Healthcare Collaborating Agency in Korea. The methodological quality was assessed by the Consensus-based Standards for the selection of health Measurement Instruments checklist box B.
Results: Ninety-five out of 12,641 studies were selected for the analysis. The median time interval for test-retest reliability was 14 days, and the ratio of sample size for test-retest reliability to the number of items in each measure ranged from 1:1 to 1:4. The most frequently used statistical methods for continuous scores was intraclass correlation coefficients (ICCs). Among the 63 studies that used ICCs, 21 studies presented models for ICC calculations and 30 studies reported 95% confidence intervals of the ICCs. Additional analyses using 17 studies that reported a strong ICC (> 0.09) showed that the mean time interval was 12.88 days and the mean ratio of the number of items to sample size was 1:5.37.
Conclusions: When researchers plan to assess the test-retest reliability of patient-reported outcome measures for older people, they need to consider an adequate time interval of approximately 13 days and the sample size of about 5 times the number of items. Particularly, statistical methods should not only be selected based on the types of scores of the patient-reported outcome measures, but should also be described clearly in the studies that report the results of test-retest reliability.

What is already known about the topic?

- Current literature has proposed common factors and quality criteria for evaluating the psychometric properties of patient-reported outcome measures.
- Of the psychometric properties, the test-retest procedure used to assess stability is exposed to several risks, such as carryover effects and actual change between two separate times.
- Although the time interval for test-retest reliability for older people might be different from that of the general population, current literature related to the quality of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported measures for older people has not been evaluated yet.

What this paper adds

- The median time interval between two administrations was 14 days.
- The mean time interval and the mean ratio of the number of items in each measure to sample size for test-retest reliability in studies that reported a strong Intraclass correlation coefficient (ICC) were

* Corresponding author at: College of Nursing Seoul National University, Daehak-ro 103, Jongro-gu, Seoul, 406-799, South Korea.
E-mail addresses: elderly1004@hanmail.net (M.S. Park), kkyungja@jejunu.ac.kr (K.J. Kang), icedcoffee@eulji.ac.kr (S.J. Jang), jylee3130@gmail.com (J.Y. Lee), changsj@snu.ac.kr (S.J. Chang).

https://doi.org/10.1016/j.ijnurstu.2017.11.003
Received 30 June 2017; Received in revised form 7 November 2017; Accepted 7 November 2017

12.88 days and 1:5.37, respectively.
- Most of studies that used continuous scores for test-retest reliability evaluated the reliability using the ICC; however, less than half these studies reported models for calculation and 95% confidence intervals of the ICC.

1. Introduction

With a great interest in concepts related to patient-centered care in the health care system, patient-reported outcomes have been emphasized around the world (Adler and Resnick, 2010). patient-reported outcomes, which indicate all kinds of information coming from patients, have been widely used for a variety purposes including evaluating health care quality, screening health risks and problems, and assessing the effects of treatment or interventions (Adler and Resnick, 2010; Deshpande et al., 2011; Nelson et al., 2015). Although the scientific knowledge about the impact of patient-reported outcomes is still debatable, the use of patient-reported outcomes in the health care system has rapidly emerged because healthcare providers can directly hear the patient's voice and obtain value from hearing the patient's perspective (Adler and Resnick, 2010; Nelson et al., 2015; Santana and Feeny, 2014).

Given the emphasis on patient-reported outcomes, attention needs be given to the measures that assess patient-reported outcomes because further plans and actions related to treatments or interventions could be changed depending on the findings of a patient-reported outcome measure (Frost et al., 2007). Hence, the psychometric properties of patient-reported outcome measures must be ensured (Deshpande et al., 2011; Frost et al., 2007; Nelson et al., 2015). Current literature has proposed common factors and quality criteria for evaluating the psychometric properties of patient-reported outcome measures; those factors are validity and reliability (Deshpande et al., 2011; Frost et al., 2007; Nelson et al., 2015; Terwee et al., 2007).

Validity is the extent to which a measure accurately measures what it is intended to measure (DeVellis, 2012; Streiner and Norman, 2008; Waltz et al., 2010). Three fundamental types of validity have been widely used to evaluate the validity of a patient-reported outcome measure: content validity refers to whether the items of a measure represent the content domain, construct validity refers to whether a measure correlates with theoretical concepts it is supposed to be related to as well as not be related to, and criterion-related validity refers to whether a measure correlates with a "gold standard" measure as a criterion (DeVellis, 2012; Streiner and Norman, 2008; Waltz et al., 2010). Reliability is the extent to which a measure is able to provide consistent and accurate results related to the target attribute (DeVellis, 2012; Polit and Beck, 2008; Waltz et al., 2010). For estimating reliability, three procedures are commonly used: internal consistency, which refers to the coherence of items within a measure; equivalence, which concerns the degree of agreement among two or more observers; and stability, which concerns obtaining comparable results at two separate times (DeVellis, 2012; Polit and Beck, 2008; Waltz et al., 2010).

Among the psychometric properties evaluating patient-reported outcome measures, the test-retest procedure used to assess stability is exposed to several risks, such as carryover effects and actual change between two separate times (DeVellis, 2012; Polit and Beck, 2008; Yu, 2005). These risks could be minimized by determining the appropriate interval between two administrations (Deshpande et al., 2011; Streiner and Norman, 2008). A short time interval might cause recall of the items, and a long time interval might permit clinical change over the time period (DeVellis, 2012; Terwee et al., 2007). Two to 14 days between the first and second administrations are generally acceptable for evaluating test-retest reliability (Streiner and Norman, 2008; Terwee et al., 2007; Waltz et al., 2010). However, the appropriate time interval could differ for diverse characteristics such as the target group's age (Frost et al., 2007; Streiner and Norman, 2008).

With rapidly emerging health issues related to aging, various patient-reported outcome measures have been developed and validated for older people. Given the changes in cognitive functioning such as memory, as well as health conditions in older people (Denton and Spencer, 2010; Gilsky, 2007), the appropriate time interval for evaluating test-retest reliability in older people might be different from that of the general population. Unfortunately, current literature related to the quality of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported outcome measures in older people has not been evaluated yet. Therefore, this study aimed to evaluate the components of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported outcome measures in older people, and to provide suggestions on the methodology for calculating test-retest reliability for patient-reported outcomes in older people.

2. Methods

This systematic review employed both the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) checklist (Equator Network, 2013) and the guideline for systematic review that was published by the National Evidence-based Healthcare Collaborating Agency (NECA) in Korea (Kim et al., 2011).

2.1. Search strategies

In order to identify the eligible studies, the electronic databases including MEDLINE, Embase, CINAHL, and PsycINFO were searched from January 1, 2000 to August 10, 2017 by an information specialist. The reason for the starting date of 2000 for the search was that the term patient-reported outcomes was suggested by the US Food and Drug Administration in 2001, although similar concepts were used from the 1970s to 1990s (Wu et al., 2013). The following search terms were determined in accordance with the PICO (population, index, comparison, and outcomes) model: population was "aged," "elderly," or "older adults"; index test was "test-retest reliability"; and comparisons and outcomes were not set for the research questions of this systematic review. The search strategies using the combined search terms in each database are delineated in Table 1. To find additional eligible studies, the researchers reviewed the reference lists of the selected studies.

Table 1
Search strategies.

	Ovid-MEDLINE	Ovid-EMBASE	CINAHL[a] complete	PsycINFO
1	Exp Aged/	Exp Aged/	(MH "Aged+")	Exp Aged/
2	elderly.mp.	elderly.mp.	(MH "Test-retest reliability")	elderly.mp.
3	"older adult$1".mp.	"older adult$1".mp.	1 AND 2	"older adult$1".mp.
4	OR/1–3	OR/1–3	limit 3 to yr = "2000–2017"	OR/1–3
5	"test-retest reliability".mp.	"test-retest reliability".mp.		"test-retest reliability".mp.
6	4 AND 5	4 AND 5		4 AND 5
7	limit 6 to yr = "2000–Current"	limit 7 to yr = "2000 − Current"		limit 7 to yr = "2000 − Current"
Results	3808	5097	3365	371

[a] CINAHL, Cumulative Index to Nursing and Allied Health Literature; The word "Current" in this table indicates between the first and second weeks in August 2017.

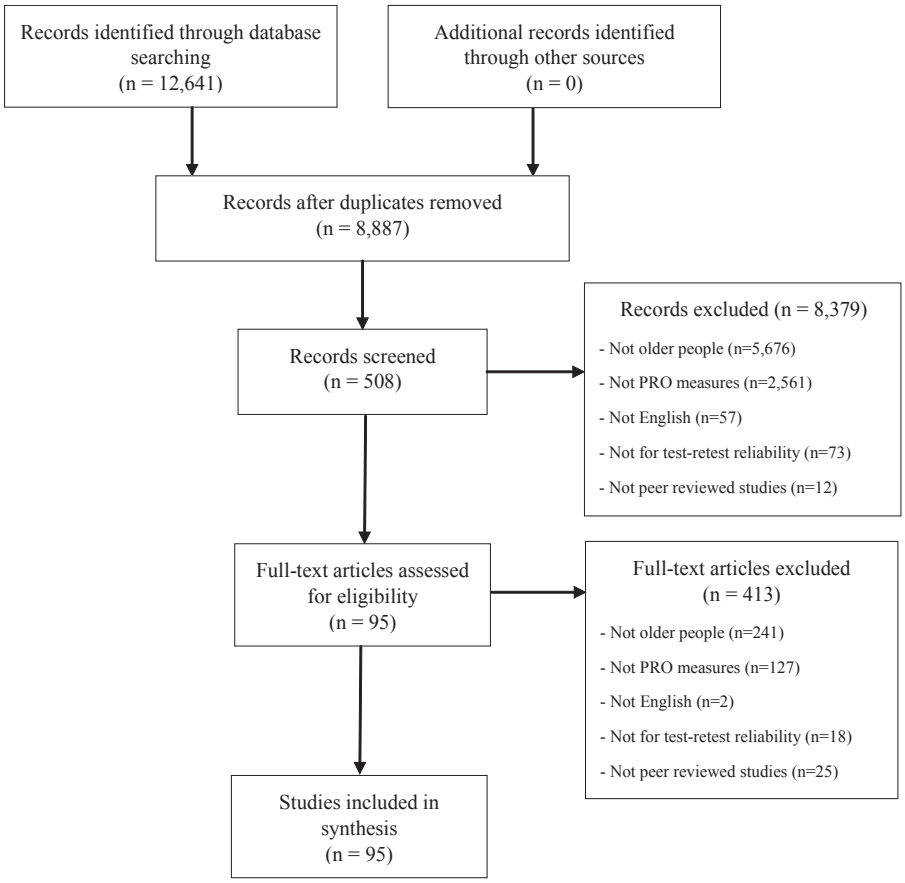

Fig. 1. Flow chart of study selection.

2.2. Inclusion and exclusion criteria

The inclusion criteria to identify the eligible studies were as follows: (a) published in English; (b) involved people aged 65 years or older; (c) evaluated test-retest reliability of patient-reported outcome measures using collected data; and (d) published in peer-reviewed journals. Studies that were excluded were (a) grey literature; (b) used measures completed by healthcare providers, caregivers, or other people; or (c) did not evaluate people aged 65 years or older (Fig. 1).

2.3. Selection of studies

Three steps as suggested by the NECA guidelines (Kim et al., 2011) were adopted to identify studies. The first step was ruling out duplicated studies among all the retrieved studies. The second step involved the initial screening by reviewing titles and abstracts of studies. The final step was assessing the full texts of the screened studies. All three steps were independently performed by four investigators. To ensure inter-rater agreement, the four investigators assessed the first 10% of the eligible studies using the inclusion criteria simultaneously (Mansutti et al., 2017). The intraclass correlation coefficient (ICC) for inter-rater agreement in the selection of studies was 0.95.

2.4. Methodological quality assessment

The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN), which was developed for assessing the methodological quality of measurement properties of patient-reported outcome measures (Mokkink et al., 2012), was used for the study. The COSMIN checklist has ten boxes related to measurement properties including internal consistency, reliability, measurement error, content validity, structural validity, hypotheses testing, cross-cultural validity, criterion validity, responsiveness, and interpretability (Mokkink et al., 2012). Among these boxes, box B concerns reliability properties and was selected for the methodological quality assessment of the selected studies in accordance with the purpose of this systematic review. Box B consists of 10 items for design requirements (missing items percentage, description related to missing items, sample size adequacy, two measurements, independent administration, time interval stated, patients' stability during the time interval, time interval adequacy, similar test conditions, and critical flaws) and four items for statistical methods (intraclass correlation for continuous scores, Kappa for dichotomous/nominal/ordinal scores, weighted Kappa for ordinal scores, and weighting scheme for ordinal scores). Each item is scored on a four-point scale rated as poor, fair, good, and excellent. The overall score of box B is determined based on the lowest score among the scores for all 14 items (Terwee et al., 2012). The methodological quality assessment in this study was independently performed by four investigators, and then the evaluation by each investigator was compared for agreement. The ICC for inter-rater agreement in the methodological quality assessment was 0.98 in this study.

2.5. Data extraction

The data extraction form was developed according to the PICO of this systematic review. From the included studies, the following data were extracted: authors; publication year; patient-reported outcome measure's name and the number of items; sample size and the mean age of the participants; sample size, time interval, statistical methods, and results for the test-retest reliability. The data extraction was conducted by four investigators who extracted data from all the included studies independently. When any disagreements emerged through comparing the extracted data, they were resolved by reviewing the study and discussing the disagreement. The ICC for inter-rater agreement in the data extraction was 0.93.

2.6. Data synthesis

For the data synthesis of this systematic review, a narrative synthesis was undertaken because the characteristics of the patient-reported outcome measures of the studies included were heterogeneous (Mays et al., 2005). The narrative synthesis was guided by the general framework for narrative synthesis (Centre for Reviews and Dissemination, 2009; Snilstveit et al., 2012). Additionally, number and percentage, median, and interquartile range (IQR, 25th to 75th percentile) were calculated in this synthesis.

In order to explore the tendency of time interval and sample size for the test-retest reliability, several steps were adopted that involved evaluating ICCs, which is a well-known standard statistical method for determining test-retest reliability for continuous scores (Hallgren, 2012; Kong, 2017). First, we chose the studies that stated the single value of ICC (e.g., 0.62 in Harada et al. (2001)'s study) among the studies that used ICC for the test-retest reliability. Then, we selected studies that had a strong ICC (> 0.9) based on Koo and Li's (2016) recommendation. Using the time interval and sample size for the test-retest reliability retrieved from these studies, the mean and 95% confidence interval (CI) were calculated.

3. Results

3.1. Search results

From the electronic databases search, a total of 12,641 studies were identified as eligible. After eliminating duplicates, 8887 studies went through the initial screening by reviewing titles and abstracts of the studies. Next, 508 studies from the initial screening were assessed by reviewing the full texts. There were no additional studies through searching by hand. Finally, a total of 95 studies met inclusion criteria and were selected for the analysis.

3.2. Methodological quality assessment results

All 95 studies were evaluated for their methodological quality using the COSMIN checklist box B. Regarding the overall methodological quality assessment, only one study by Roaldsen et al. (2014) was rated as good. The methodological quality of 73 studies were fair, while 21 studies were evaluated as poor.

Of all items on the COSMIN checklist box B, item 4, which related to at least two measurements, was rated as excellent in all 95 studies. Regarding item 6 that related to time interval being stated, only one study (Ngai et al., 2012) that did not provide the time interval was rated as fair. Seventy-nine studies were excellent on item 10, which related to critical flaw of study design. On the other hand, the item that was most frequently rated as poor was item 3, which related to sample size adequacy. This was because 14 out of the 95 studies used a sample size below 30 for test-retest reliability. The overall methodological quality assessment results are included in Table 2.

3.3. Description of the studies

The general characteristics of the studies are summarized in Table 2. Among the 95 studies, the total sample size ranged from 20 (Cohen-Mansfield and Jensen, 2007) to 1200 (Hwang et al., 2003) (median 147, IQR 71.5 to 288).

Regarding the patient-reported outcome measures, the most frequently used concept was activity (28.4%), followed by quality of life (15.8%) and fall (11.6%). Additionally, health care satisfaction (3.2%), depression (3.2%), and self-care (3.2%) were used as a core concept of patient-reported outcome measures. The number of items in each measure ranged from one (McCormack et al., 2011; Gill et al., 2012) to 115 (Cohen-Mansfield and Jensen, 2007) (median 20, IQR 12 to 29.5).

3.4. Description on test-retest reliability

Test-retest reliability was analyzed by focusing on time interval, sample size, and statistical methods for the test-retest reliability (Tables 2 and 3).

3.4.1. Time interval

With regard to time interval between two separate administrations, 39 (41.1%) studies applied the second administration between 14 and 20 days after the first one. Thirteen (13.6%) studies had a time interval greater than 28 days, and one study (Ngai et al., 2012) did not report time interval for the test-retest reliability. The time interval reported in the studies ranged from 2 to 3 days (Ferrell et al., 2000) to 632.9 days (Heisel and Flett, 2016). Interestingly, Adler and Resnick (2010) administered the Dementia Quality of Life measure at 4- and 12-months after initial testing.

Among the 95 studies, 78 (82.1%) suggested a single (e.g., 14 days) or average (e.g., 5.6 days) time interval, whereas 17.9% stated the range of time interval (e.g., from 21 to 28 days). The median time interval of 78 studies was 14 days (IQR 7 to 14).

3.4.2. Sample size

Thirty (31.6%) studies administered the test-retest to 30 to 49 older persons, and 22 (23.1%) studies had more than 100 older persons complete the test-retest. The range of sample size for the test-retest reliability was from 10 (Komjakraphan et al., 2009; Robichaud and Lamarre, 2002; Huang, 2006) to 663 (Jiang et al., 2017) (median 44, IQR 30 to 81.5).

Regarding the ratio of the number of items in each measure to sample size for evaluating test-retest reliability, the ratio of 38 (40.0%) studies was from 1:1 to 1:4. The ratio ranged from 1:0.1 (Robichaud and Lamarre, 2002) to 1:47.4 (Jiang et al., 2017) (median 1:2.6, IQR 1:1.4 to 1:6.9).

3.4.3. Statistical methods

Among the 95 studies, 23 (24.2%) used two statistical methods for evaluating test-retest reliability. For example, Boele van Hensbroek et al. (2009) applied the ICC for the total scores and Kappa coefficient for the individual item score of the patient-reported outcome measure. Muntinga et al. (2014) evaluated test-retest reliability using both the ICC and standard measurement errors. A study on Portuguese elders used both the ICC and Pearson correlation coefficient (Pocinho et al., 2009). Hence, a total of 118 statistical methods in 95 studies were used, and the most frequently used statistical methods for test-retest reliability were ICCs (53.4%), followed by correlation coefficients (23.7%).

Of the 95 studies, 88 that used continuous scores evaluated test-retest reliability by using the ICC (n = 63, 71.5%), correlation coefficients (n = 24, 27.2%), or Kappa coefficients (n = 1, 1.3%). Seven studies that used categorical scores such as dichotomous, nominal, and ordinal scores applied Kappa or weighted Kappa coefficients for analyzing test-retest reliability (Ettema et al., 2007; Gill et al., 2012; Levy, 2003; McCormack et al., 2011; Muller et al., 2016; Seo et al., 2017; Stathokostas et al., 2012).

Twenty-one out of 63 studies that used ICC for evaluating test-retest reliability provided the ICC model, such as a two-way random effects model (Muntinga et al., 2014) and the ICC (2.1) (Newell et al., 2012). Thirty of 63 studies reported 95% CIs of the ICC (Koo and Li, 2016).

3.5. Tendency of time interval and sample size for test-retest reliability with strong ICC

For this analysis, 35 out of 63 studies that used ICC were chosen because they reported the single value of ICC. Among them, 17 studies that had a strong ICC (greater than 0.9) based on Koo and Li's (2016) recommendation were selected.

As depicted in Table 4, the mean time interval between two

Table 2
Characteristics of the included study (n = 95).

Patient-reported Outcomes measures			First Author, year	Sample size	Mean age (years)	Test-retest reliability				QA[a]
Concept	Measure name	No of items				Sample size	Time interval (days)	Statistical methods	Results	
Activity	Adherence to Exercise Scale for Older Patients	43	Hardage (2007)	50	79.9	28	14	ICC[d]	ICC = 0.328–0.796	Fair
	Community Healthy Activities Model Program for Seniors	84	Harada (2001)	87	75.0	80	14	Correlation(P) [c]	r = 0.62	Fair
								ICC[d]	ICC = 0.62	
	Community Healthy Activities Model Program for Seniors (A)	A: 9	Gennuso (2016)	58	75.1	58	10	ICC[d]	ICC = 0.64(A)	Fair
	Yale Physical Activity Survey for Older Adults (B)	B: 1							ICC = 0.59(B)	
	Engagement in Meaningful Activities survey	12	Eakman (2010)	154	80.5	25	7–14	Correlation(P)[c]	r = 0.56	Fair
	Elderly EXERNET Physical Activity Questionnaire	3	López-Rodríguez (2017)	73	71.9	73	14	ICC[d]	ICC = 0.68–0.97	Fair
	Geriatric Locomotive Function Scale	25	Seichi (2012)	711	77	205	14	ICC[d]	ICC = 0.712–0.924	Fair
	Incidental and Planned exercise questionnaire(IPEQ)-WA (A)	10	Delbaere (2010)	500	77.4	A:50	7	ICC[d]	ICC = 0.84(A)	Fair
	IPEQ-W (B)					B:50			ICC = 0.77(B)	
	International Physical Activity Questionnaire	19	Van Holle (2015)	434	74.2	29	9.6	ICC[d]	ICC = 0.43–0.81	Poor
	Keele Assessment of Participation	11	Hermsen (2013)	364	76.5	122	14	ICC[d]	ICC = 0.57–0.63	Fair
	Late-Life Function and Disability Instrument	48	Elboim-Gabyzeon (2015)	61	74.1	41	6–8	ICC[d]	ICC = 0.77–0.93	Fair
								SEM	SEM = 4.6–7	
	Late-Life Function and Disability Instrument-computer adaptive	20	Arensman (2016)	54	80.2	54	2–8 (mean = 5)	ICC[d]	ICC = 0.76–.81	Poor
	Modified Gait Efficacy Scale	10	Newell (2011)	102	78.6	26	28–31	ICC[d]	ICC = 0.93	Poor
	Neighborhood Environment Walkability Scale	76	Cerin (2010)	484	≧65	92	14–20 (mean = 17)	ICC[d]	ICC = 0.52–0.77	Fair
								Percent agreement (PA)	PA = 0.38–0.99	
	PaArticular Scales	86	Muller (2016)	191	80.6	40	3	Unweighted Kappa Weighted Kappa	Kappa(U) = 0.67 Kappa(W) = 0.78	Fair
	PaArticular Scales	86	Muller (2016)	43	79.4	43	7	ICC[d]	ICC = 0.74–0.88	Fair
	PaArticular Scales	86	Muller (2016)	159	80	43	7	Weighted Kappa	Kappa = 0.75(absolute PA) Kappa = 0.56(relative PA)	Fair
	Physical Activity Scale for the Elderly	12	Hagiwara (2008)	325	72.6	257	21–28	ICC[d]	ICC = 0.65	Fair
	Physical Activity Scale for the Elderly	12	Vaughan (2013)	73	79	66	14	ICC[d] (total) Kappa(item)	ICC = 0.79 Kappa = 0.17–0.92	Fair
	Physical Activity Scale for the Elderly	12	Ayvat (2017)	80	69.5	80	7	ICC[d]	ICC = 0.99	Fair
	Physical Activity Scale for the Elderly	12	Ngai (2012)	90	77.7	32	not stated	ICC[d]	ICC = 0.81	Fair
	Tilburg Frailty Indicator	25	Coelho (2015)	252	79.2	74	mean = 14	Correlation (P)[c](total) Kappa(item)	r = 0.91 Kappa = 0.52–0.95	Fair
	Vulnerable Elders Survey-13	13	Carneiro (2015)	206	73.0	22	1–30	Correlation (P)[c](total) Kappa(item)	r = 0.92 Kappa = 0.62–1.00	Poor
	Questionnaire for exercise-related injury	12	Stathokostas (2012)	110	72.3	110	7	Kappa	Kappa = 0.76–1.00	Fair
	Scale of Older Adult's Routine	42	Zisberg (2009)	80	84.7	78	14	ICC[d] (continuous score) Kappa(nominal score)	ICC = 0.46–0.91 Kappa > 0.75	Fair

(*continued on next page*)

Table 2 (*continued*)

Patient-reported Outcomes measures			First Author, year	Sample size	Mean age (years)	Test-retest reliability				QA[a]
Concept	Measure name	No of items				Sample size	Time interval (days)	Statistical methods	Results	
	Self-Assessment of Physical Fitness scale	3	Weening-Dijksterhuis (2012)	76	86.0	76	7	ICC[d]	ICC = 0.66-0.70	Fair
	Self-Efficacy for Exercise Scale	9	Rydwik (2014)	39	72.0	34	mean = 17	ICC[d] (total) Weighted Kappa (item)	ICC = 0.79 Kappa = 0.63	Fair
	Tampa Scale for Kinesiophobia	11	Larsson (2014)	433	74.8	262	14	ICC[d] (total) Weighted Kappa (item)	ICC = 0.747 Kappa = 0.366–0.579	Fair
	Valued Activity Inventory for Adults with Cancer	25	Lyons (2012)	50	75.0	40	3	ICC[d]	ICC = 0.65	Fair
QoL[b]	12-item Short-Form Health Survey	12	Resnick (2001)	182	86.0	182	14–28	Correlation(P)[c]	r = 0.73–86	Fair
	15D instrument of health-related quality of life	15	Okamoto (2013)	423	74.0	423	28	ICC[d]	ICC = 0.44–0.72	Fair
	36-item Short-Form Health Survey	36	Han (2004)	219	73.7	52	21–25	Correlation(P)[c]	r = 0.710–0.895	Fair
	COOP/WONCA[g]	6	Ettema (2007)	67	84.6	34	7	Weighted Kappa	Kappa = 0.23–.67	Fair
	Dementia Quality of life	29	Wolak (2010)	109	81.1	109	14	ICC[d]	ICC = 0.96–1.00	Fair
	Dementia Quality of life	29	Adler (2010)	486	83.8	231	120 365	Correlation[c]	r = 0.60(at 120-day) r = 0.635(at 365-day)	Fair
	ICECAP-O[h](A) ASCOT[i](B) EQ-5D-3L[j](C)	A:5 B:8 C:5	Van Leeuwen (2015)	190	82.4	147	7–14	ICC[d] SEM	ICC > 0.70 SEM = not stated	Fair
	LEIPAD	49	Jalenques (2013)	195	72.6	50	15	ICC[d]	ICC = 0.70–0.90	Fair
	Mealtime Satisfaction Questionnaire	15	Pizzola (2013)	147	88.0	64	21	ICC[d] (total) Kappa(item)	ICC = 0.91 Kappa = 0.33–0.64	Fair
	Oral impacts on daily performance	10	Nair (2016)	202	75.0	202	30	ICC[d]	ICC = 0.75	Fair
	Psychological well-being instrument	9	Ottenbacher (2007)	40	76.3	40	mean = 5	ICC[d]	ICC =0.66–0.79	Fair
	Sarcopenia Quality of Life	55	Beaudart (2017)	43	77.1	30	14	ICC[d]	ICC = 0.91	Fair
	WHOQOL-BREF[e]	26	Hwang (2003)	1200	73.4	30	60	ICC[d]	ICC = 0.59–0.94	Fair
	WHOQOL-BREF[e]	26	Naumann (2004)	39	78.1	39	7	ICC[d]	ICC = 0.919	Fair
	WHOQOL-OLD[f]	24	Peel (2007)	74	≥65	44–53	14	Correlation(S)[c]	r = 0.51–0.95, ICC = 0.57–0.96	Fair
	SF-12	12						ICC[d]		
Fall	Activities Specific Balance Confidence Scale	16	Ayhan (2014)	106	69.5	106	14	ICC[d]	ICC = 0.997	Fair
	Activities-Specific Balance Confidence	16	Lallawmzuali (2017)	100	≥65	25	14	ICC[d]	ICC = 98	Poor
	CAREFALL Triage-Instrument	44	Boele van Hensbroek (2009)	300	73.0	27	14	ICC[d] (total)	ICC = 0.79	Poor
								Kappa(item)	Kappa = 0.34–0.78	
	Falls Efficacy Scale	12	Büla (2008)	70	81.1	20	3	ICC[d]	ICC = 0.97	Poor
	Falls Efficacy Scale International	16	Ulus (2012)	70	69.7	70	14	ICC[d]	ICC = 0.97–0.99(item) ICC = 0.94(total)	Fair
	Falls Efficacy Scale-International (FES-I) (A)	A:16	Kempen (2008)	213	76.6	193	28	ICC[d]	ICC = 0.82 (A)	Fair
	Short FES-I (B)	B:7							ICC = 0.83 (B)	
	Falls Efficacy Scale-International (FES-I) (A)	A:16	Ruggiero (2009)	157	79.4	151	28	Correlation(S)[c]	r = 0.88 (A)	Fair
	Short FES-I (B)	B:7							r = 0.87 (B)	
	Geriatric Fear of Falling Measure	41	Huang (2006)	100	75.4	10	14	Correlation(P)[c]	r = 0.88	Fair
	Late-Life Function and Disability Instrument	48	Roaldsen (2014)	62	76.0	62	14	ICC[d]	ICC = 0.82–0.91	Good
								SEM	SEM = 2.9–5.1	
	Measure of Balance Confidence	10	Simpson (2009)	45	81.0	45	7	ICC[d]	ICC = 0.96	Fair

(*continued on next page*)

Table 2 (*continued*)

Patient-reported Outcomes measures			First Author, year	Sample size	Mean age (years)	Test-retest reliability				QA[a]
Concept	Measure name	No of items				Sample size	Time interval (days)	Statistical methods	Results	
	Self-assessment of falls risk	20	Elliott (2004)	52	81.0	52	2–14 (mean = 5.6)	Correlation(P)[c] (total) Kappa(item)	r = 0.91 Kappa = 0.36–0.91	Fair
Satisfaction	CareWell in Hospital patient questionnaire	8	Bakker (2014)	470	76.9	78	2–14	ICC[d] (total) Kappa(item)	ICC = 0.75 Kappa = 0.28–0.82	Fair
	Health Care Satisfaction Questionnaire	23	Gagnon (2006)	873	82.0	38	16	ICC[d]	ICC = 0.72	Fair
	Satisfaction with the Nursing Home Instrument	29	Lee (2006)	330	81.5	330	14	ICC[d]	ICC = 0.94	Fair
Depression	Brief Assessment Schedule Depression Cards (A)	A:19	Healey (2008)	49	78.8	43	7–10	Correlation(K)[c]	τ = 0.66 (A)	Poor
	Beck Depression Inventory-Fast Screen (B)	B:7							τ = 0.63 (B)	
	Geriatric Depression Scale	27	Pocinho (2009)	200	76.6	38	8	Correlation(P)[c] ICC[d]	r = 0.995 ICC = 0.979	Fair
	Single-item depression screener	1	McCormack (2011)	65	79.0	20	7	Kappa	Kappa = 0.43	Fair
Pain	Geriatric Pain Measure	24	Ferrell (2000)	176	84.7	50	2–3	Correlation(P)[c] Kappa	r = 0.90 Kappa = 0.60	Fair
	Geriatric Pain Measure	24	Park (2009)	121	69.1	32	14–28	Correlation(P)[c]	r = 0.643	Fair
Self-efficacy	Diabetes Self-Efficacy Scale	8	Chang (2014)	278	75.2	14	28	ICC[d] Percent agreement (PA)	ICC = 0.80 PA = not stated	Poor
	Chronic Disease Self-Efficacy Scales	6	Chow (2014)	163	75.3	12	14	ICC[d]	ICC = 0.98	Poor
Self-care	Diabetes Self-Management Behavior for Older Koreans	14	Seo (2017)	150	76.5	22	14	Kappa Percent agreement (PA)	Kappa = −0.07 to 1.0 PA = 0.32–1.0	Poor
	Geriatric Health Promotion scale	22	Wang (2015)	520	75.0	520	180	Correlation[c]	r = 0.72	Fair
	Health-promoting lifestyles profile	36	Cao (2012)	1012	70.8	21	7	Correlation[c]	r = 0.68	Fair
	Medication-Risk Questionnaire	10	Levy, 2003	40	73.0	20	7	Kappa	Kappa > 0.6	Fair
	Self-maintenance Habits and Preferences in Elderly	115	Cohen-Mansfield (2007)	20	79.8, 80.4	20	7–14	ICC[d]	ICC = 0.72	Poor
Disease-related	Cancer Knowledge scale	29	Su (2009)	214	75.0	31	14	Correlation(P)[c]	r = 0.83	Fair
	Patient Generated Index	6	Tully (2002)	479	≥65	236	14	ICC[d]	ICC = 0.39–0.58	Fair
	Western Ontario and McMaster Index(WOMAC®)	24	Papathanasiou (2015)	123	69.5	123	7	ICC[d]	ICC = 0.95	Fair
Anxiety	Geriatric Anxiety Inventory	20	Pachana (2007)	46	78.8	46	7	Correlation(P)[c]	r = 0.91	Fair
	Geriatric Anxiety Inventory	20	Kneebone (2016)	81	79.0	63	median = 7	Correlation(K)[c]	τ = 0.53	Poor
Attitude	Life Attitude Scale	27	Liu (2001)	663	≥65	37	14	Correlation(P)[c]	r = 0.87∼0.96	Fair
Client-centered	Client-Centered Care Questionnaire	15	Muntinga (2014)	389	83.0	67	7–14	ICC[d] SEM	ICC = 0.81 SEM = 2.61	Fair
Cognitive impairment	Patient-Reported Outcomes in Cognitive Impairment	55	Frank (2006)	186	77.0	43 (23%)	14	ICC[d]	ICC = 0.49–0.90	Fair
Communication	Geriatric Index of Communicative Ability	31	Kim (2014)	100	69.4	30	14	Correlation(P)[c]	r = 0.58–0.98	Poor
Coping	The Inventory of Coping Strategies Used by the Elderly	94	Robichaud (2002)	64	78.8	10	14	ICC[d]	ICC = 0.83	Fair
Driving	Driving Comfort Scales (A) Perceived Driving Abilities (B) Situational Driving Frequency (C) Situational Driving Avoidance (D)	A:13 B:15 C:14 D:20	Blanchard (2010)	61	80.4	39	mean = 7.6	ICC[d]	ICC = 0.89–0.92(A) ICC = 0.65–0.66(B) ICC = 0.89(C) ICC = 0.86(D)	Fair
	Safe Driving Behavior Measure	68	Song (2016)	61	68.6	61	3	ICC[d]	ICC = 0.92	Poor
Dyspnea	Dyspnea Management Questionnaire	30	Norweg (2006)	85	76.0	26	mean = 17.9	ICC[d]	ICC = 0.71–0.95	Poor

(*continued on next page*)

Table 2 (*continued*)

Patient-reported Outcomes measures			First Author, year	Sample size	Mean age (years)	Test-retest reliability				QA[a]
Concept	Measure name	No of items				Sample size	Time interval (days)	Statistical methods	Results	
Empower-ment	Empowerment Questionnaire for Inpatients scale	16	Lopez (2010)	87	73.7	28	7–10	Correlation[c] (total) Kappa(item)	r = 0.88 Kappa = 0.10–0.93	Poor
Hearing	Self-assessment for Hearing Screening of the Elderly	20	Kim (2016)	83	77.3	83	21	Correlation(P)[c]	r = 0.76-0.85	Fair
Life-space utility	Life-Space Assessment Questionnaire	9	Kammerlind (2014)	298	80.0	298	14	ICC[d] (total) Weighted Kappa	ICC = 0.84 Kappa = 0.50–0.94	Poor
Motivation	Motivations for Living Inventory	15	Wang (2013)	247	81.7	40	14	ICC[d]	ICC = 0.81	Fair
Perception	Aging Perceptions Questionnaire	32	Chen (2016)	94	71.8	30	14	ICC[d]	ICC = 0.82–0.10	Fair
Reminis-cence	Modified Reminiscence Functions Scale	39	Washington (2009)	271	≥65	32	14	Correlation(P)[c]	r = 0.82	Poor
Spiritual	Daily Spiritual Experience Scale	6	Bailly (2010)	338	77.9	40	14	Correlation(P)[c]	r = 0.85	Fair
	Spirituality Index of Well-Being	12	Wu (2017)	416	81.1	416	14	ICC[d]	ICC = 0.99	Fair
Stress	Perceived Stress Scale	14	Jiang (2017)	663	79.2	663	365	ICC[d]	ICC = 0.62	Fair
Suicide	Geriatric Suicide Ideation Scale	66	Heisel (2006)	107	81.5	32	52	Correlation(P)[c]	r = 0.86	Poor
	Geriatric Suicide Ideation Scale	31	Heisel (2016)	173	73.9	A:146 B:126 C:112	A:mean = 29.5 B:mean = 399.5 C:mean = 632.9	ICC[d] Correlation(P)[c]	ICC = 0.80, 0.77, 0.64 r = 0.70, 0.59, 0.49	Fair
	Protective Reasons against Suicide Inventory	20	Wang (2016)	200	≥65	30	28	ICC[d]	ICC = 0.95	Fair
Support	Thai Family Support Scale for Elderly Parents	61	Komjakraphan (2009)	500	70.7 (n = 60)	10	7	Kappa	Kappa = 0.96	Poor

[a] QA, quality assessment.
[b] QoL, quality of life.
[c] Correlation(P = Pearson, S = Spearman, K = Kendall).
[d] ICC, intraclass correlation.
[e] WHOQOL-BREF, Brief version of the World Health Organization Quality of Life.
[f] WHOQOL-OLD, World Health Organization Quality of Life for older adults.
[g] COOP/WONCA, Dartmouth Cooperative Functional Assessment Charts/World Organization of General Practitioners/Family Physicians.
[h] ICECAP-O, ICEpop CAPability measure for older people.
[i] ASCOT, Adult Social Care Outcomes Toolkit.
[j] EQ-5D-3L, EuroQol five-dimensional questionnaire.

Table 3
Characteristics of test-retest reliability.

Characteristics	Time interval (n = 95)		Sample size (n = 95)				Statistical methods (n = 118)[a]	
	Days	N (%)	Number of participants	N (%)	Ratio of items to sample size	N (%)	Types	N (%)
	< 7	8 (8.4)	≤29	20 (21.1)	< 1:1	18 (18.9)	ICC[b]	63 (53.4)
	7≤ ≤13	28 (29.5)	30≤ ≤49	30 (31.6)	1:1≤ < 1:4	38 (40.0)	Correlations[c]	28 (23.7)
	14≤ ≤20	39 (41.1)	50≤ ≤99	23 (24.2)	1:4≤ < 1:8	17 (17.9)	Kappa[d]	20 (16.9)
	21≤ ≤27	6 (6.3)	100≤	22 (23.1)	1:8≤ < 1:12	6 (6.3)	SEM[e]	4 (3.5)
	28≤	13 (13.6)	–		1:12≤	16 (16.8)	PA[f]	3 (2.5)
	Not stated	1 (1.1)	–		–		–	
Range	3 to 365[g]		10 to 663		1:0.1 to 1:47.4		–	
Median	14[g]		44		1:2.6			
IQR[h]	7 to 14[g]		30 to 81.5		1:1.4 to 1:6.9			

[a] It is because some studies used two statistical methods for the test-retest reliability.
[b] ICC, intraclass correlation coefficients.
[c] Correlations include Pearson and Spearman correlation coefficients.
[d] Kappa includes Cohen's Kappa and weighted Kappa.
[e] SEM, standard error of measurement.
[f] PA, percent agreement.
[g] The range, median, and IQR of time interval was calculated based on the studies which were suggested single value or average value of time interval (n = 78).
[h] IQR, interquartile range.

administrations for the 17 studies was 12.88 days (95% CI 8.91 to 16.85 days). Regarding the ratio of the number of items in each measure to sample size, the mean ratio was 1:5.37 with a 95% CI of 1:1.23 to 1:9.51.

4. Discussion

This systematic review analyzed studies that evaluated the test-retest reliability of patient-reported outcome measures for older people. Throughout this review, we observed several interesting findings that we will discuss in accordance with components of test-retest reliability, including time interval, sample size, and statistical methods used for test-retest reliability.

For time interval, we found that the majority of the studies had a time interval from seven to 20 days between the two administrations of the test. This interval is slightly longer than the suggested interval of two to 14 days for test-retest reliability (Streiner and Norman, 2008; Terwee et al., 2007; Waltz et al., 2010). The differences might be caused by the categorization of time interval in this systematic review. For this reason, we performed further analysis by using the data from studies that had a single or average time interval. Interestingly, about half of the studies used 14 days for the time interval, so the values of the median and 75th percentile were the same 14 days. Based on the analyses, we found that considerable studies that aimed to validate the test-retest reliability for older people used approximately 14 days. This time interval also is supported by the finding that studies that reported an ICC of 0.9, which is considered strong, on average used 12.88 days for the test-retest. As such, the time interval for older people in this review was consistent with general recommendations that suggest one to two weeks (Streiner and Norman, 2008; Terwee et al., 2007). Hence, the findings on the time interval for test-retest reliability in this review indicated that there might be no special time interval for older people (e.g., shorter time interval). However, it should be prudently translated because this review did not take into account older people's diseases that could affect cognitive functioning.

Table 4
Tendency of time interval and sample size for test-retest reliability with strong ICC (n = 17).

ICC[†]	Time interval (days)		Ratio of items to sample size	
	Mean	95% CI[‡](lower, upper)	Mean	95% CI[‡](lower, upper)
> 0.9 (Strong)	12.88	(8.91, 16.85)	1:5.37	(1:1.23, 1:9.51)

[†] ICC, intraclass correlation coefficients.
[‡] CI, confidence interval.

With respect to sample size for test-retest reliability, about half of the studies had 30 to 99 participants, indicating fair to good methodological quality in accordance with the COSMIN checklist box B; in which less than 30 participants, 30 to 49 participants, 50 to 99 participants, and greater than 100 participants denote poor, fair, good, and excellent methodological quality, respectively (Terwee et al., 2012). However, this category of sample size for test-rest reliability was proposed based on a rule of thumb (Stevens, 1996), while the total sample size for psychometric evaluation of a measure is calculated depending on the number of items (e.g., 10-fold per item). Thus, an additional analysis was conducted to explore the ratio of sample size to the number of items on the patient-reported outcome measures. We found the majority of studies used from one- to eight-fold per item, and this finding was also supported by the results from the analysis of studies that reported a strong ICC.

When it comes to the statistical methods used to evaluate test-retest reliability, almost all studies selected adequate statistical methods based on the ICC being recommended for continuous scores and the Kappa is regarded as a suitable method for categorical scores (Hallgren, 2012; Kong, 2017; Terwee et al., 2007). However, the use of correlation coefficients for evaluating test-retest reliability is debatable (Streiner and Norman, 2008). This is because the nature of reliability reflects two facets degree of correlation and agreement (Bruton et al., 2000). The focus of correlation analyses to explain the relationship rather than the agreement among variables; hence, the correlation coefficient could be at risk of reporting greater than the pure reliability (Koo and Li, 2016; Streiner and Norman, 2008).

Interestingly, several studies applied two statistical methods including the ICC for the total scores (a continuous variable) and Kappa coefficients for the individual item score (categorical variables) (Bakker et al., 2014; Larsson et al., 2014; Boele van Hensbroek et al., 2009).

The finding that most studies used continuous scores for test-retest reliability and the majority of them used ICC as a statistical method is consistent with a previous study that reported the quality of test-retest reliability of measures on symptoms and health-related quality of life in cancer patients (Paiva et al., 2014). It might be that researchers prefer ICC because it considers systematic changes between two measurements as well as it can be calculated using the data from small sample sizes (Lexell and Downham, 2005; Vaz et al., 2013).

Regarding the reporting of ICC, about 30% of the studies that used ICC reported the model for the ICC calculation and approximately 50%

reported the 95% CIs. Even though the number of studies that reported the ICC's model has increased compared with the finding in the study by Paiva et al. (2014) in which no studies described the model for the ICC calculation, most of the studies that used ICC still provided a single value of ICC. ICC could be calculated by 10 different ways that result in different findings and interpretations, and the ICC method using a 2-way mixed-effects model for test-retest reliability is recommended because of non-randomized samples (Portney and Watkins, 2000). Using different models for the ICC calculation could result in different ICC values from the same data set (Koo and Li, 2016). Hence, the ICC model should be described in research reports. In addition, given the ICC's 95% CIs help the researcher determine the true ICC value, those should be provided in the research report.

Some limitations that should be acknowledged for the interpretation of this systematic review are described as follows. First, the psychometric properties, including test-retest reliability, reflect data collected in a particular study and not the disposition of the patient reported measures (Beckstead, 2013). Therefore, when interpreting the findings of this systematic review, this should be considered. Second, this systematic review did not consider characteristics of patient-reported outcome measures such as the type of scale (e.g., dichotomous, Likert) and framework (e.g., norm-referenced measures, which assess relative levels compared to a well-defined group, and criterion-referenced measures, which assess relative levels compared to an established goal of target behaviors) (Waltz et al., 2010). Even though the assessment methods of test-retest reliability are similar for different frameworks, the method of scoring should be considered because the reliability coefficients could be affected by it (Waltz et al., 2010). Although there were several studies that validated patient-reported outcome measures having the same concepts such as physical activity and quality of life, the patient-reported outcome measures almost always differed from each other. Because of this heterogeneity, a meta-analysis could not be conducted. Finally, due to limitations on resources we decided to only include peer-reviewed studies published in English and we did not conduct searches for grey literature in conference proceedings, web searches, and studies that were published in languages other than English.

5. Conclusions

This systematic review concludes with some implications. When researchers plan to validate the test-retest reliability of patient-reported outcome measures for older people, they should consider an adequate time interval, the sample size, and statistical methods. Particularly, statistical methods should not only be selected based on the types of scores of the patient-reported outcome measures, but should also be described clearly in the studies that report the results of test-retest reliability. Additionally, the characteristics of patient-reported outcome measures should be considered in future studies analyzing the quality of test-retest reliability, including time interval, sample size, and statistical methods.

Disclosure statement

No competing financial interests exist.

Acknowledgement

The study was supported by 2017 SNU invitation program for distinguished scholar of Seoul National University.

References

Adler, E., Resnick, B., 2010. Reliability and validity of the Dementia Quality of Life measure in nursing home residents. West. J. Nurs. Res. 32, 686–704. http://dx.doi.org/10.1177/0193945909360780.

Arensman, R.M., Pisters, M.F., de Man-van Ginkel, J.M., Schuurmans, M.J., Jette, A.M., de Bie, R.A., 2016. Translation, validation, and reliability of the Dutch late-life function and disability instrument computer adaptive test. Phys. Ther. 96, 1430–1437. http://dx.doi.org/10.2522/ptj.20150265.

Ayhan, C., Büyükturan, Ö., Kirdi, N., Yakut, Y., Güler, Ç., 2014. The Turkish version of the activities specific balance confidence (abc) scale: its cultural adaptation, validation and reliability in older adults. Turk. J. Geriatr. 17, 157–163.

Ayvat, E., Kilinç, M., Kirdi, N., 2017. The Turkish version of the Physical Activity Scale for the Elderly (PASE): its cultural adaptation, validation, and reliability. Turk. J. Med. Sci. 47, 908–915. http://dx.doi.org/10.3906/sag-1605-7.

Bailly, N., Roussiau, N., 2010. The Daily Spiritual Experience Scale (DSES): validation of the short form in an elderly French population. Can. J. Aging. 29, 223–231. http://dx.doi.org/10.1017/S0714980810000152.

Bakker, F.C., Persoon, A., Schoon, Y., Olde Rikkert, M.G., 2014. The CareWell in Hospital questionnaire: a measure of frail elderly inpatient experiences with individualized and integrated hospital care. J. Hosp. Med. 9, 324–329. http://dx.doi.org/10.1002/jhm.2158.

Beaudart, C., Biver, E., Reginster, J.Y., Rizzoli, R., Rolland, Y., Bautmans, I., Petermans, J., Gillain, S., Buckinx, F., Dardenne, K., Bruyère, O., 2017. Validation of the SarQoL, a specific health-related quality of life questionnaire for Sarcopenia. 8, 238–244. https://doi.org/10.1002/jcsm.12149.

Beckstead, J.W., 2013. On measurement and their quality: paper 1: reliability – history, issues and procedures. Int. J. Nurs. Stud. 50, 968–973. http://dx.doi.org/10.1016/j.ijnurstu.2013.04.005.

Blanchard, R.A., Myers, A.M., 2010. Examination of driving comfort and self-regulatory practices in older adults using in-vehicle devices to assess natural driving patterns. Accid. Anal Prev. 42, 1213–1219. http://dx.doi.org/10.1016/j.aap.2010.01.013.

Boele van Hensbroek, P., van Dijk, N., van Breda, G.F., Scheffer, A.C., van der Cammen, T.J., Lips, P., Goslings, J.C., de Rooij, S.E., Combined Amsterdam and Rotterdam Evaluation of FALLs (CAREFALL) study group, 2009. The CAREFALL Triage instrument identifying risk factors for recurrent falls in elderly patients. Am. J. Emerg. Med. 27, 23–36. http://dx.doi.org/10.1016/j.ajem.2008.01.029.

Bruton, A., Conway, J.H., Holgate, S.T., 2000. Reliability: what is it, and how is it measured? Physiotherapy 86, 94–99. http://dx.doi.org/10.1016/S0031-9406(05)61211-4.

Büla, C.J., Martin, E., Rochat, S., Piot-Ziegler, C., 2008. Validation of an adapted falls efficacy scale in older rehabilitation patients. Arch. Phys. Med. Rehabil. 89, 291–296. http://dx.doi.org/10.1016/j.apmr.2007.08.152.

Cao, W.J., Chen, C.S., Hua, Y., Li, Y.M., Xu, Y.Y., Hua, Q.Z., 2012. Factor analysis of a health-promoting lifestyle profile (HPLP): application to older adults in Mainland China. Arch. Gerontol. Geriatr. 55, 632–638. http://dx.doi.org/10.1016/j.archger.2012.07.003.

Carneiro, F., Sousa, N., Azevedo, L.F., Saliba, D., 2015. Vulnerability in elderly patients with gastrointestinal cancer – translation, cultural adaptation and validation of the European Portuguese version of the Vulnerable Elders Survey (VES-13). BMC Cancer 15 (723). http://dx.doi.org/10.1186/s12885-015-1739-2.

Centre for Reviews and Dissemination, 2009. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. https://www.google.co.kr/?gfe_rd=cr&ei=v8ZQWZvZFc3o8AeZtrHwCw&gws_rd=ssl#q=+Systematic+reviews:+CRD's+guidance+for+undertaking+reviews+in+health+care.+&spf=1498467181233 (Accessed 14.03.01).

Cerin, E., Sit, C.H., Cheung, M.C., Ho, S.Y., Lee, L.C., Chan, W.M., 2010. Reliable and valid NEWS for Chinese seniors: measuring perceived neighborhood attributes related to walking. Int. J. Behav. Nutr. Phys. Act. 25 (84). http://dx.doi.org/10.1186/1479-5868-7-84.

Chang, S.J., Song, M., Im, E.O., 2014. Psychometric evaluation of the Korean version of the diabetes self-efficacy scale among South Korean older adults with type 2 diabetes. J. Clin. Nurs. 23, 2121–2130. http://dx.doi.org/10.1111/jocn.12133.

Chen, X., Hu, Y., Zhu, D., Li, J., Zhou, L., 2016. Chinese version of the Aging Perceptions Questionnaire (C-APQ): assessment of reliability and validity. Aging. Ment. Health. 20, 567–574. http://dx.doi.org/10.1080/13607863.2015.1028332.

Chow, S.K., Wong, F.K., 2014. The reliability and validity of the Chinese version of the short-form Chronic disease self-efficacy scales for older adults. J. Clin. Nurs. 23, 1095–1104. http://dx.doi.org/10.1111/jocn.12298.

Coelho, T., Santos, R., Paul, C., Gobbens, R.J., Fernandes, L., 2015. Portuguese version of the Tilburg Frailty Indicator: Transcultural adaptation and psychometric validation. 15, 951–960. https://doi.org/10.1111/ggi.12373.

Cohen-Mansfield, J., Jensen, B., 2007. Self-maintenance Habits and Preferences in Elderly (SHAPE): reliability of reports of self-care preferences in older persons. Aging Clin. Exp. Res. 19, 61–68.

Delbaere, K., Hauer, K., Lord, S.R., 2010. Evaluation of the Incidental and Planned Exercise Questionnaire (IPEQ) for older people. Br. J. Sports Med. 44, 1029–1034. http://dx.doi.org/10.1136/bjsm.2009.060350.

Denton, F.T., Spencer, B.G., 2010. Chronic health conditions: changing prevalence in an aging population and some implications for the delivery of health care services. Can. J. Aging 29, 11–21. http://dx.doi.org/10.1017/S0714980809990390.

Deshpande, P.R., Rajan, S., Sudeepthi, B.L., Abdul Nazir, C.P., 2011. Patient-reported outcomes: a new era in clinical research. Perspect. Clin. Res. 2, 137–144. http://dx.doi.org/10.4103/2229-3485.86879.

DeVellis, R.F., 2012. Scale Development: Theory and Applications. SAGE Publications, Inc, Thousand Oaks, California.

Eakman, A.M., Carlson, M., Clark, F., 2010. Factor structure, reliability, and convergent validity of the engagement in meaningful activities survey for older adults. OTJR 30, 111–121.

Elboim-Gabyzon, M., Agmon, M., Azaiza, F., Laufer, Y., 2015. Translation and validation of the Arab version of the late-life function and disability instrument: a cross sectional

study. BMC Geriatr. 15, 51. http://dx.doi.org/10.1186/s12877-015-0046-8.
Elliott, J.A., Jamieson, J.L., Donnelly, M.L., Malone, M., 2004. Measurement properties of a new falls risk self-assessment questionnaire for seniors. J. Can. Geratr. Soc. 7, 98–102.
Ettema, T.P., Hensen, E., De Lange, J., Droes, R.M., Mellenbergh, G.J., Ribbe, M.W., 2007. Self report on quality of life in dementia with modified COOP/WONCA charts. Aging Ment. Health 11, 734–742. http://dx.doi.org/10.1080/13607860701366236.
Equator Network, 2013. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. . http://www.equator-network.org/reporting-guidelines/prisma/ (Accessed 14.03.01).
Ferrell, B.A., Stein, W.M., Beck, J.C., 2000. The geriatric pain measure: validity, reliability and factor analysis. J. Am. Geriatr. Soc. 48, 1669–1673.
Frank, L., Flynn, J.A., Kleinman, L., Margolis, M.K., Matza, L.S., Beck, C., Bowman, L., 2006. Validation of a new symptom impact questionnaire for mild to moderate cognitive impairment. Int. Psychogeriatr. 18, 135–149.
Frost, M.H., Reeve, B.B., Liepa, A.M., Stauffer, J.W., Hays, R.D., Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group, 2007. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 10 (2), S94–S105.
Gagnon, M., Hébert, R., Dubé, M., Dubois, M.F., 2006. Development and validation of the Health Care Satisfaction Questionnaire (HCSQ) in elders. J. Nurs. Meas. 14, 190–204.
Gennuso, K.P., Matthews, C.E., Colbert, L.H., 2015. Reliability and validity of 2 self-report measures to assess sedentary behavior in older adults. J. Phys. Act. Health 12, 727–732. http://dx.doi.org/10.1123/jpah.2013-0546.
Gill, D.P., Jones, G.R., Zou, G., Speechley, M., 2012. Using a single question to assess physical activity in older adults: a reliability and validity study. BMC Med. Res. Methodol. 12http://dx.doi.org/10.1186/1471-2288-12-20. 20-2288-12-20.
Gilsky, E.L., 2007. Changes in cognitive function in human aging. In: Riddle, D.R. (Ed.), Brain Aging: Models, Methods, and Mechanisms. CRC Press, Boca Raton, Florida, pp. 3–20.
Hagiwara, A., Ito, N., Sawai, K., Kazuma, K., 2008. Validity and reliability of the Physical Activity Scale for the Elderly (PASE) in Japanese elderly people. Geriatr. Gerontol. Int. 8, 143–151. http://dx.doi.org/10.1111/j.1447-0594.2008.00463.x.
Harada, N.D., Chiu, V., King, A.C., Stewart, A.L., 2001. An evaluation of three self-report physical activity instruments for older adults. Med. Sci. Sports Exerc. 33, 962–970.
Hardage, J., Peel, C., Morris, D., Graham, C., Brown, C., Foushee, H.R., Braswell, J., 2007. Adherence to Exercise Scale for Older Patients (AESOP): a measure for predicting exercise adherence in older adults after discharge from home health physical therapy. J. Geriatric. Phys. Ther. 30, 69–78.
Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. 8, 23–34.
Han, C.W., Lee, E.J., Iwaya, T., Kataoka, H., Kohzuki, M., 2004. Development of the Korean version of short-form 36-item health survey: health related QOL of healthy elderly people and elderly patients in Korea. Tohoku J. Exp. Med. 203, 189–194.
Healey, A.K., Kneebone, I.I., Carroll, M., Anderson, S.J., 2008. A preliminary investigation of the reliability and validity of the brief assessment schedule depression cards and the beck depression inventory-fast screen to screen for depression in older stroke survivors. Int. J. Geriatr. Psychiatry 23, 531–536.
Heisel, M.J., Flett, G.L., 2006. The development and initial validation of the geriatric suicide ideation scale. Am. J. Geriatr. Psychiatry 14, 742–751.
Heisel, M.J., Flett, G.L., 2016. Investigating the psychometric properties of the Geriatric Suicide Ideation Scale (GSIS) among community-residing older adults. Aging Ment. Health 20, 208–221. http://dx.doi.org/10.1080/13607863.2015.1072798.
Hermsen, L.A., Terwee, C.B., Leone, S.S., van der Zwaard, B., Smalbrugge, M., Dekker, J., van der Horst, H.E., Wilkie, R., 2013. Social participation in older adults with joint pain and comorbidity; testing the measurement properties of the Dutch Keele Assessment of Participation. BMI Open. http://dx.doi.org/10.1136/bmjopen-2013-003181.
Huang, T.T., 2006. Geriatric fear of falling measure: development and psychometric testing. Int. J. Nurs. Stud. 43, 357–365. http://dx.doi.org/10.1016/j.ijnurstu.2005.04.006.
Hwang, H.F., Liang, W.M., Chiu, Y.N., Lin, M.R., 2003. Suitability of the WHOQOL-BREF for community-dwelling older people in Taiwan. Age Ageing 32, 593–600.
Jalenques, I., Auclair, C., Roblin, J., Morand, D., Tourtauchaux, R., May, R., Vaille-Perret, E., Watts, J., Gerbaud, L., De Leo, D., 2013. Cross-cultural evaluation of the French version of the LEIPAD, a health-related quality of life instrument for use in the elderly living at home. Qual. Life. Res. 22, 509–520. http://dx.doi.org/10.1007/s11136-012-0166-y.
Jiang, J.M., Seng, E.K., Zimmerman, M.E., Sliwinski, M., Kim, M., Lipton, R.B., 2017. Evaluation of the reliability, validity, and predictive validity of the subscales of the perceived stress scale in older adults. J. Alzheimers Dis. 59, 987–996. http://dx.doi.org/10.3233/JAD-170289.
Kammerlind, A.S., Fristedt, S., Bravell, M.E., Fransson, E.I., 2014. Test-retest reliability of the Swedish version of the life-space assessment questionnaire among community-dwelling older adults. Clin. Rehabil. 28, 817–823. http://dx.doi.org/10.1177/0269215514522134.
Kempen, G.I., Yardley, L., van Haastregt, J.C., Zijlstra, G.A., Beyer, N., Hauer, K., Todd, C., 2008. The Short FES-I: a shortened version of the falls efficacy scale-international to assess fear of falling. Age Ageing 37, 45–50.
Kim, G., Na, W., Kim, G., Han, W., Kim, J., 2016. The development and standardization of self-assessment for hearing screening of the elderly. Clin. Interv. Aging 16, 787–795. http://dx.doi.org/10.2147/CIA.S107102.
Kim, J.W., Nam, C.M., Kim, Y.W., Kim, H.H., 2014. The development of the Geriatric Index of Communicative Ability (GICA) for measuring communicative competence of elderly: a pilot study. Speech Commun. 56, 63–69. http://dx.doi.org/10.1016/j.specom.2013.08.001.
Kim, S.Y., Park, J.E., Seo, H.J., Jang, B.H., Son, H.J., Sun, H.S., Shin, C.M., 2011. NECA's Guidance for Undertaking Systematic Reviews and Meta-analyses for Intervention. http://www.neca.re.kr/center/researcher/book_view.jsp?boardNo=CA&seq=5819&q=626f6172644e6f3d4341 (accessed 16.09.01).
Kneebone, I.I., Fife-Schaw, C., Lincoln, N.B., Harder, H., 2016. A study of the validity and the reliability of the geriatric anxiety inventory in screening for anxiety after stroke in older inpatients. Clin. Rehabil. 30, 1220–1228. http://dx.doi.org/10.1177/0269215515619661.
Komjakraphan, R., Isalamalai, S., Boonyasopun, U., Schneider, J.K., 2009. Development of the thai family support scale for elderly parents (TFSS-EP). Thai J. Nurs. Res. 13, 118–132.
Kong, K.A., 2017. Statistical methods: reliability assessment and method comparison. Ewha Med. J. 40, 9–16. https://doi.org/10.12771/emj.2017.40.1.9.
Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15, 155–163. http://dx.doi.org/10.1016/j.jcm.2016.02.012.
Lallawmzuali, R.S., Neha, G., 2017. Validation of the Mizo version of the Activities-specific Balance Confidence (ABC) scale. Indian J. Physiother. Occup. Ther. 11, 112–117. http://dx.doi.org/10.5958/0973-5674.2017.00022.3.
Larsson, C., Hansson, E.E., Sundquist, K., Jakobsson, U., 2014. Psychometric properties of the Tampa Scale of Kinesiophobia (TSK-11) among older people with chronic pain. Physiother. Theory Pract. 30, 421–428. http://dx.doi.org/10.3109/09593985.2013.877546.
Lee, L.Y., Lee, D.T., Woo, J., Wong, E.M., 2006. Validation of the Chinese version of the satisfaction with the nursing home instrument. J. Clin. Nurs. 15, 1574–1582.
Levy, H.B., 2003. Self-administered medication-risk questionnaire in an elderly population. Ann. Pharmacother. 37, 982–987. http://dx.doi.org/10.1345/aph.1C305.
Lexell, J.E., Downham, D.Y., 2005. How to assess the reliability of measurements in rehabilitation. Am. J. Phys. Med. Rehabil. 84, 719–723.
Liu, S.J., 2001. The construction and evaluation of the reliability and validity of a life attitude scale for elderly with chronic disease. J. Nurs. Res. 9, 33–42.
Lopez, J.E., Orrell, M., Morgan, L., Warner, J., 2010. Empowerment in older psychiatric inpatients: development of the empowerment questionnaire for inpatients (EQuIP). Am. J. Geriatr. Psychiatry 18, 21–32. http://dx.doi.org/10.1097/JGP.0b013e3181b2090b.
López-Rodríguez, C., Laquna, M., Gomez-Cabello, A., Gusi, N., Espino, L., Villa, G., Pedrero-Chamizo, R., Cadajus, J.A., Ara, I., Azna, S., 2017. Validation of the self-report EXERNET questionnaire for measuring physical activity and sedentary behavior in elderly. Arch. Gerontol. Geriatr. 69, 156–161. http://dx.doi.org/10.1016/j.archger.2016.11.004.
Lyons, K.D., Hegel, M.T., Hull, J.G., Balan, S., Bartels, S., 2012. Reliability and validity of the valued activity inventory for adults with cancer. OTJR 32, 238–245. http://dx.doi.org/10.3928/15394492-20110623-02.
Mansutti, I., Saiani, L., Grassetti, L., Palese, A., 2017. Instruments evaluating the quality of the clinical learning environment in nursing education: a systematic review of psychometric properties. Int. J. Nurs. Stud. 68, 60–72. http://dx.doi.org/10.1016/j.ijnurstu.2017.01.001.
Mays, N., Pople, C., Popay, J., 2005. Details of Approaches to Synthesis a Methodological Appendix to the Paper: Systematically Reviewing Qualitative and Quantitative Evidence to Inform Management and Policy Making in the Health Field. https://pdfs.semanticscholar.org/92fd/665a527a1d3e6cf6b751725ffb2d3b2b4fed.pdf (Accessed 16.09.01).
McCormack, B., Boldy, D., Lewin, G., McCormack, G.R., 2011. Screening for depression among older adults referred to home care services: a single-item depression screener versus the geriatric depression scale. Home Health Care Manage. Pract. 23 (1), 13–19.
Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C., 2012. COSMIN Checklist Manual. http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf (Accessed 16.09.01).
Muller, M., Oberhauser, C., Fischer, U., Bartoszek, G., Saal, S., Strobl, R., Meyer, G., Grill, E., 2016. The Particular Scales: a new outcome measure to quantify the impact of joint contractures on activities and participation in individuals in geriatric care: development and Rasch analysis. Int. J. Nurs. Stud. 59, 107–117. http://dx.doi.org/10.1016/j.ijnurstu.2016.04.002.
Muntinga, M.E., Mokkink, L.B., Knol, D.L., Nijpels, G., Jansen, A.P., 2014. Measurement properties of the Client-centered Care Questionnaire (CCCQ): factor structure, reliability and validity of a questionnaire to assess self-reported client-centeredness of home care services in a population of frail, older people. Qual. Life Res. 23, 2063–2072. http://dx.doi.org/10.1007/s11136-014-0650-7.
Nair, R., Tsakos, G., Yee Ting Fai, R., 2016. Testing reliability and validity of oral impacts on daily performances for Chinese-speaking elderly Singaporeans. Gerodontology 33, 499–505. http://dx.doi.org/10.1111/ger.12192.
Naumann, V.J., Byrne, G.J., 2004. WHOQOL-BREF as a measure of quality of life in older patients with depression. Int. Psychogeriatr. 16, 159–173.
Nelson, E.C., Eftimovska, E., Lind, C., Hager, A., Wasson, J.H., Lindblad, S., 2015. Patient reported outcome measures in practice. BMJ 350, g7818. http://dx.doi.org/10.1136/bmj.g7818.
Newell, A.M., VanSwearingen, J.M., Hile, E., Brach, J.S., 2012. The modified Gait Efficacy Scale: establishing the psychometric properties in older adults. Phys. Ther. 92, 318–328. http://dx.doi.org/10.2522/ptj20110053.
Norweg, A.M., Whiteson, J., Demetis, S., Rey, M., 2006. A new functional status outcome measure of dyspnea and anxiety for adults with lung disease: the dyspnea management questionnaire. J. Cardiopulm. Rehabil. 26, 395–404.
Ngai, S.P.C., Cheung, R.T.H., Lam, P.L., Chiu, J.K.W., Fung, E.Y.H., 2012. Validation and reliability of the physical activity scale for the elderly in Chinese population. J. Rehabil. Med. 44, 462–465. http://dx.doi.org/10.2340/16501977-0953.

Okamoto, N., Hisashige, A., Tanaka, Y., Kurumatani, N., 2013. Development of the Japanese 15D instrument of health-related quality of life: verification of reliability and validity among elderly people. PLoS One 8, e61721. http://dx.doi.org/10.1371/journal.pone.0061721.
Ottenbacher, M.E., Kuo, Y.F., Ostir, G.V., 2007. Test-retest reliability of a psychological well-being scale in hospitalized older adults. Aging. Clin. Exp. 16, 424–429.
Pachana, N.A., Byrne, G.J., Siddle, H., Koloski, N., Harley, E., Arnold, E., 2007. Development and validation of the Geriatric Anxiety Inventory. Int. Psychogeriatr. 19, 103–114.
Paiva, C.E., Barroso, E.M., Carneseca, E.C., de Padua Souza, C., Dos Santos, F.T., Mendoza Lopez, R.V., Ribeiro Paiva, S.B., 2014. A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. BMC Med. Res. Methodol. 14, 8. http://dx.doi.org/10.1186/1471-2288-14-8.
Papathanasiou, G., Stasi, S., Oikonomou, L., Roussou, I., Papageorgiou, E., Chronopoulos, E., Korres, N., Bellamy, N., 2015. Clinimetric properties of WOMAC Index in Greek knee osteoarthritis patients: comparisons with both self-reported and physical performance measures. Rheumatol. Int. 35, 115–123. http://dx.doi.org/10.1007/s00296-014-3043-x.
Park, J., Cho, B., Paek, Y., Kwon, H., Yoo, S., 2009. Development of a pain assessment tool for the older adults in Korea: the validity and reliability of a Korean version of the geriatric pain measure (GPM-K). Arch. Gerontol. Geriatr. 49, 199–203. http://dx.doi.org/10.1016/j.archger.2008.07.010.
Peel, N.M., Bartlett, H.P., Marshall, A.L., 2007. Measuring quality of life in older people: reliability and validity of WHOQOL-OLD. Australas. J. Ageing 26, 162–167. http://dx.doi.org/10.1111/j.1741-6612.2007.00249.x.
Pizzola, L., Martos, Z., Pfisterer, K., de Groot, L., Keller, H., 2013. Construct validation and test–retest reliability of a mealtime satisfaction questionnaire for retirement home residents. J. Nutr. Gerontol. Geriatr. 32, 343–359. http://dx.doi.org/10.1080/21551197.2013.840257.
Pocinho, M.T.S., Farate, C., Dias, C.A., Lee, T.T., Yesavage, J.A., 2009. Clinical and psychometric validation of the Geriatric Depression Scale (GDS) for Portuguese elders. J. Clin. Geron 32, 223–236. http://dx.doi.org/10.1080/07317110802678680.
Polit, D.F., Beck, C.T. (Eds.), 2008. Nursing Research: Generating and Assessing Evidence for Nursing Practice. Lippincott Williams & Wilkins Philadelphia, Pennsylvania.
Portney, L.G., Watkins, M.P., 2000. Foundations of Clinical Research: Applications to Practice. Prentice Hall, New Jersey.
Resnick, B., Parker, R., 2001. Simplified scoring and psychometrics of the revised 12-item short-form health survey. Outcomes Manag. Nurs. Pract. 5, 161–166.
Roaldsen, K.S., Halvarsson, A., Sarlija, B., Franzen, E., Stahle, A., 2014. Self-reported function and disability in late life – cross-cultural adaptation and validation of the Swedish version of the late-life function and disability instrument. Disabil. Rehabil. 36, 813–817. http://dx.doi.org/10.3109/09638288.2013.819387.
Robichaud, L., Lamarre, C., 2002. Developing an instrument for identifying coping strategies used by the elderly to remain autonomous. Am. J. Phys. Med. Rehabil. 81, 736–744. http://dx.doi.org/10.1097/01.CCM.0000026923.24522.0F.
Rydwik, E., Hovmöller, F., Boström, C., 2014. Aspects of reliability and validity of the Swedish version of the Self-Efficacy for Exercise Scale for older people. Physiother. Theory Pract. 30, 131–137. http://dx.doi.org/10.3109/09593985.2013.838614.
Ruggiero, C., Mariani, T., Gugliotta, R., Gasperini, B., Patacchini, F., Nguyen, H.N., Zampi, E., Serra, R., Dell'Aquila, G., Cirinei, E., Cenni, S., Lattanzio, F., Cherubini, A., 2009. Validation of the Italian version of the falls efficacy scale international (FES-I) and the short FES-I in community-dwelling older persons. Arch. Gerontol. Geriatr. 49 (Suppl. 1), 211–219. http://dx.doi.org/10.1016/j.archger.2009.09.031.
Santana, M.J., Feeny, D., 2014. Framework to assess the effects of using patient-reported outcome measures in chronic care management. Qual. Life Res. 23, 1505–1513. http://dx.doi.org/10.1007/s11136-013-0596-1.
Seichi, A., Hoshino, Y., Doi, T., Akai, M., Tobimatsu, Y., Iwaya, T., 2012. Development of a screening tool for risk of locomotive syndrome in the elderly: the 25-question geriatric locomotive function Scale. J. Orthop. Sci. 17, 163–172. http://dx.doi.org/10.1007/s00776-011-0793-5.
Seo, K., Song, M., Choi, S., Kim, S.A., Chang, S.J., 2017. Development of a scale to measure diabetes self-management behaviors among older Koreans with type 2 diabetes, based on the seven domains identified by the American association of diabetes educators. Jpn. J. Nurs. Sci. 14, 161–170. http://dx.doi.org/10.1111/jjns.12145.
Simpson, J.M., Worsfold, C., Fisher, K.D., Valentine, J.D., 2009. The CONFbal scale: a measure of balance confidence–a key outcome of rehabilitation. Physiotherapy 95, 103–109. http://dx.doi.org/10.1016/j.physio.2008.12.004.
Song, C.S., Lee, J.H., Han, S.W., 2016. Test-retest reliability of the safe driving behavior measure for community-dwelling elderly drivers. J. Phys. Ther. Sci. 28, 1716–1719. http://dx.doi.org/10.1589/jpts.28.1716.
Snilstveit, B., Oliver, S., Vojtkova, M., 2012. Narrative approaches to systematic review and synthesis of evidence for international development policy and practice. J. Dev. Effect. 4, 409–429. http://dx.doi.org/10.1080/19439342.2012.710641.
Stathokostas, L., Theou, O., Vandervoort, T., Raina, P., 2012. Psychometric properties of a questionnaire to assess exercise-related musculoskeletal injuries in older adults attending a community-based fitness facility. BMJ Open 2, e001777. http://dx.doi.org/10.1136/bmjopen-2012-001777.
Stevens, J., 1996. Applied Multivariate Statistics for the Social Science. Lawrence Erlbaum, Mahway, New Jersey.
Streiner, D.L., Norman, G.R., 2008. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford University Press, New York.
Su, C.C., Chen, Y.M., Kuo, B.J., 2009. Development and psychometric testing of the cancer knowledge scale for elders. J. Clin. Nurs. 18, 700–707. http://dx.doi.org/10.1111/j.1365-2702.2008.02489.x.
Terwee, C.B., Bot, S.D., de Boer, M.R., van der Windt, D.A., Knol, D.L., Dekker, J., Bouter, L.M., de Vet, H.C., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60, 34–42. http://dx.doi.org/10.1016/j.jclinepi.2006.03.012.
Terwee, C.B., Mokkink, L.B., Knol, D.L., Ostelo, R.W., Bouter, L.M., de Vet, H.C., 2012. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual. Life Res. 21, 651–657. http://dx.doi.org/10.1007/s11136-011-9960-1.
Tully, M., Cantrill, J., 2002. The test-retest reliability of the modified patient generated index. J. Health Serv. Res. Policy 7, 81–89.
Ulus, Y., Durmus, D., Akyol, Y., Terzi, Y., Bilgici, A., Kuru, O., 2012. Reliability and validity of the Turkish version of the Falls Efficacy Scale International (FES-I) in community-dwelling older persons. Arch. Gerontol. Geriatr. 54, 429–433. http://dx.doi.org/10.1016/j.archger.2011.06.010.
Van Holle, V., De Bourdeaudhuil, I., Deforche, B., Van Cauwenberg, J., Van Dyck, D., 2015. Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L). BMC Public Health 15, 433. http://dx.doi.org/10.1186/s12889-015-1785-3.
Van Leeuwen, K.M., Bosmans, J.E., Janse, A.P., Hoogendijk, E.O., van Tulder, M.W., van der Horst, H.E., Ostelo, R.W., 2015. Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. Value Health 18, 35–43. http://dx.doi.org/10.1016/j.jval.2014.09.006.
Vaughan, K., Miller, W.C., 2013. Validity and reliability of the Chinese translation of the Physical Activity Scale for the Elderly (PASE). Disabil. Rehabil. 35, 191–197. http://dx.doi.org/10.3109/09638288.2012.690498.
Vaz, S., Falkmer, T., Passmore, A.E., Parsons, R., Andreou, P., 2013. The case for using the repeatability coefficient when calculating test-retest reliability. PLoS One 8, e73990. https://doi.org/10.1371.journal.pone.0073990.
Weening-Dijksterhuis, E., de Greef, M.H., Krijnen, W., van der Schans, C.P., 2012. Self-reported physical fitness in frail older persons: reliability and validity of the Self-Assessment of Physical Fitness (SAPF). Percept. Mot. Skills 115, 797–810.
Waltz, C.F., Strickland, O.L., Lenz, E.R., 2010. Measurement in Nursing and Health Research. Springer Publishing Company, New York, NY.
Wang, J., Lee, C.M., Chang, C.F., Jane, S.W., Chen, M.Y., 2015. The development and psychometric testing of the geriatric health promotion scale. J. Nurs. Res. 23, 56–64. http://dx.doi.org/10.1097/jnr.0000000000000077.
Wang, Y.W., Tsai, Y.F., Lee, S.H., Chen, Y.J., Chen, H.F., 2016. Development and psychometric testing of the protective reasons against suicide inventory for assessing older Chinese-speaking outpatients in primary care settings. J. Adv. Nurs. 72, 1701–1710. http://dx.doi.org/10.1111/jan.12971.
Wang, Y.W., Tsai, Y.F., Wong, T.K., Ku, Y.C., 2013. Development and psychometric testing of a Chinese-language instrument for assessing institutionalised older males' motivations for living. J. Clin. Nurs. 22, 2867–2875. http://dx.doi.org/10.1111/jocn.12089.
Washington, G., 2009. Modification and psychometric testing of the reminiscence functions scale. J. Nurs. Meas. 17, 134–147.
Wolak, A., Jolly, D., Dramé, M., Boyer, F., Morrone, I., Aquino, J.P., Rouaud, O., Perret Guillaume, C., Ravenel, E., Dantoine, T., Ankri, J., Blanchard, F., Novella, J.L., 2010. Quality of life in dementia: psychometric properties of a French language version of the Dementia Quality of Life questionnaire (DQoL). Eur. Geriatr. Med. 1, 334–347. http://dx.doi.org/10.1016/j.eurger.2010.09.008.
Wu, A.W., Kharrazi, H., Boulware, L.E., Snyder, C.F., 2013. Measure once, cut twice–adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. J. Clin. Epidemiol. 66, S12–S20. https://doi.org/j.jclinepi/2013.04.005.
Wu, L.F., Yang, S.H., Koo, M., 2017. Psychometric properties of the Chinese version of spiritual index of well-being in elderly Taiwanese. BMC. Geriatr. 17, 3. http://dx.doi.org/10.1186/s12877-016-0392-1.
Yu, C.H., 2005. Test-retest reliability. In: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Academic Press, San Diego CA, pp. 777–784.
Zisberg, A., Young, H.M., Schepp, K., 2009. Development and psychometric testing of the scale of older adults' routine. J. Adv. Nurs. 65, 672–683. http://dx.doi.org/10.1111/j.1365-2648.2008.04901.x.