OPEN

Toward Patient-Centered Care: A Systematic Review of How to Ask Questions That Matter to Patients

Alicia Rosenzveig, BSc, Ayse Kuspinar, MSc, Stella S. Daskalopoulou, MD, PhD, and Nancy E. Mayo, PhD

Abstract: Clinicians rarely systematically document how their patients are feeling. Single item questions have been created to help obtain and monitor patient relevant outcomes, a requirement of patient-centered care.

The objective of this review was to identify the psychometric properties for single items related to health aspects that only the patient can report (health perception, stress, pain, fatigue, depression, anxiety, and sleep quality). A secondary objective was to create a bank of valid single items in a format suitable for use in clinical practice.

Data sources used were Ovid MEDLINE (1948 to May 2013), EMBASE (1960 to May 2013), and the Cumulative Index to Nursing and Allied Health Literature (1960 to May 2013).

For the study appraisal, 24 articles were systematically reviewed. A critical appraisal tool was used to determine the quality of articles.

Items were included if they were tested as single items, related to the construct, measured symptom severity, and referred to recent experiences.

The psychometric properties of each item were extracted. Validity and reliability was observed for the items when compared with clinical interviews or well-validated measures. The items for general health perception and anxiety showed weak to moderate strength correlations ($r = 0.28$–0.70). The depression and stress items showed good area under the receiver operating characteristic curve of 0.85 and 0.73–0.88, respectively, with high sensitivity and specificity. The fatigue item demonstrated a strong effect size and correlations up to $r = 0.80$. The 2 pain items and the sleep item showed high reliability (intraclass correlation coefficient [ICC] $= 0.85$, $\kappa = 0.76$, ICC $= 0.9$, respectively).

The search targeted articles about psychometric properties of single items. Articles that did not have this as the primary objective may have been missed. Furthermore, not all the articles included had the complete set of psychometric properties for each item.

There is sufficient evidence to warrant the use of single items in clinical practice. They are simple, easily implemented, and efficient and thus provide an alternative to multi-item questionnaires. To facilitate their use, the top performing items were combined into the visual analog health states, which provides a quick profile of how the patient is feeling. This information would be useful for regular long-term monitoring.

(*Medicine* 93(22):e120)

Abbreviations: ClinRO = clinician-reported outcome, ES = effect size, FAS = face anxiety scale, GRS = graphic rating scale, HRQL = health-related quality of life, ICC = intraclass correlation coefficient, ObsRO = observer-reported outcome, PRO = patient-reported outcome, PROMIS = patient-reported outcomes measurement information system, ROC = receiver operating characteristic, VAHS = visual analog health state, VAMS = visual analog mood scale, VAS = visual analog scale, VRS = verbal rating scale.

INTRODUCTION

Forming a collaborative relationship between the patient and the clinician is the cornerstone of patient-centered care. This dialogue must focus on patients' concerns, which cannot always be inferred from the clinical diagnosis and may need to be elicited through direct questioning. The literature suggests that patient-centered concerns, apart from survival, are symptoms, function, and health-related quality of life (HRQL).[1]

Although a profile of clinical health status requires comprehensive and systematic assessments of organ and system function, a profile of patient-centered outcomes would involve the administration of multiple items or questionnaires, the results from which need to be interpreted, monitored, and reinterpreted over time. Clinicians take a systematic approach to assessing clinical health status but are not commonly systematic in obtaining information on symptoms, functions, and HRQL—outcomes that matter to patients. There are a number of measures of HRQL, which have been used to improve quality of care in clinical practice[2] with disappointing results.[3–10] HRQL measures can provide tremendous insight into the matters of concern to patients and can track within-patient changes over time.[11] Despite these benefits, HRQL measures are not yet widely used in patient care.[12] One reason may be because clinicians face an uncertainty as to what the HRQL scores mean and how to apply the information. There is a lack of clarity about specific use of the information, including the ability to screen for problems, monitor progress over time, and facilitate models of patient-centered care.[13] For this reason, a number of single items or questions have been developed to streamline the obtaining of this important information from patients.[14–17] Collections are advantageous over multi-item questionnaires because they are simple and easy to implement,[18] quick,[18] less cognitively demanding,[19] and can be stored as a reference to compare the score over time.[18,20] To be useful in clinical practice, the single items

Editor: Suzanne Morse Buhrow.
Received: May 15, 2014; revised and accepted: August 16, 2014.

From the Division of Clinical Epidemiology (AR, NEM), McGill University Health Centre; School of Physical and Occupational Therapy (AK, NEM); and Department of Medicine (SSD), Faculty of Medicine, McGill University, Montreal, QC, Canada.
Correspondence: Nancy E. Mayo, Division of Clinical Epidemiology, McGill University Health Centre, 687 Avenue des Pins Ouest, Montreal, QC, H3A 1A1, Canada (e-mail: nancy.mayo@mcgill.ca).

The authors have no funding and conflicts of interest to disclose.

ISSN: 0025-7974
DOI: 10.1097/MD.0000000000000120

must first relate strongly to the construct they have been developed to represent, as well as have the psychometric properties important for any measure. Thus, convergent or criterion validity (sensitivity and specificity if criterion is binary) is primordial for these single items, and evidence of reliability and responsiveness are additionally necessary.

Single-item patient-centered health outcome measures include specific performance tests (eg, walking speed) and asking questions to the patient directly. When questions relate to information on outcomes that only the patient can provide, these are termed patient-reported outcomes (PROs). PROs have been defined as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else,"[21] whereas non-PROs can be clinician-reported outcomes (ClinROs), observer-reported outcomes (ObsRO), or tests of performance (PerfROs). ClinROs arise from the results of a physical examination; the Apgar and the Preschool Respiratory Assessment Measure[22] are examples. ObsROs are measures based on an observer (such as a family caregiver) assessing the patient's behavior; examples include the Dementia Rating Scale,[23] the brain impairment behavior scale,[24] and the emotional behavior index form.[25] The six-minute walk test,[26] and the Barthel Index[27] are examples of PerfROs, although the Barthel Index can also be completed by self-report or as an ObsRO. Although both ClinROs, ObsROs, and PerfROs can provide a physician with the status of a patient, there is no better way to improve patient-centered care than to increase the usage of PROs and ask patients directly about outcomes that matter to them.

The specific objective of this review was to identify the psychometric properties for single-PRO items through a search of the literature. A secondary objective was to create a bank of valid single items in a suitable format for clinical practice.

METHODS

Domains Under Study in Systematic Review

A systematic review was carried out. Specific domains were chosen based on PROs meaningful for improved patient-centered care and that were not assessable through physical examination or performance testing. Seven domains were selected as meeting these criteria, also because they are often queried directly or indirectly within a health care encounter: general health perception, stress, pain, fatigue, depression, anxiety, and sleep.

Search Methods for Identification of Studies

Articles were identified by searching the following databases: Ovid MEDLINE (1948 to May 2013), EMBASE (1960 to May 2013), and the Cumulative Index to Nursing and Allied Health Literature (1960 to May 2013). The same principle was used to search each database for each domain, which included the following terms: depression/mood, fatigue/energy, anxiety, sleep/sleep quality, pain, stress/distress, and self-rated health/general health perception. Each of these terms were combined with the following terms: validity OR concurrent validity OR discriminant validity OR construct validity OR criterion-related validity OR validation studies OR instrument validation OR known-groups OR discriminant analysis OR reliability OR intrarater reliability OR test-retest reliability OR interrater reliability OR intraclass correlation coefficient OR kappa statistic OR minimally significant OR minimal detectable OR psychometrics OR responsiveness OR minimally important change OR meaningful change OR minimal detectable OR minimally significant. Finally, all of these search terms were combined with the term single-item using the Boolean operator AND.

This literature review was supplemented by known articles of valid single items: the visual analog mood scales (VAMS)[28] and the global items in the patient-reported outcomes measurement information system (PROMIS).[29]

As this was a literature review and did not involve recruitment of patients, ethical approval was not necessary.

Article Screening and Data Extraction

Only full publications in peer-reviewed journals were considered. Unpublished data, abstracts, grey literature, and studies published in languages other than English or French were excluded. All study designs (randomized controlled trials, cross-sectional studies, etc.) and health conditions were included. The study population was restricted to adults.

Two authors (A.R. and A.K.) independently screened the citations and abstracts identified in the search and amalgamated them into Reference Manager 12, Thomson Reuters. Based on the abstracts, duplicate or irrelevant articles were excluded, as were full text articles that did not meet inclusion criteria.

Item Selection and Inclusion Criteria

Items were included in the systematic review based on the following criteria:

1. developed as a single- or stand-alone item or had been tested for this intent even if it had originally been part of a multi-item measure;
2. appeared to relate strongly to the construct it had been developed to represent;
3. measured severity, not impact; and
4. referred to recent experience (current or past week).

The multidimensionality of measuring symptoms can create challenges as they can vary in severity, duration, frequency, and impact.[30] We therefore excluded items referring to the impact because it can change as people modify their activities and lifestyles,[30] reporting low impact because of curtailment of activities without there being actual change in symptom intensity, duration, or frequency. Thus, it was also important to choose items where the wording directly tapped the construct rather than its consequences. For example, an item referring to depression needs to use words referring to elements of mood and not to the degree of engagement in activities or roles. Additionally, we favored items assessing the current time frame and not those requiring historical averaging (eg, past 2 weeks, past month). Ultimately, items were selected if deemed useful in a clinical setting, were quick to administer, and yielded answers meaningful for both clinicians and patients.

Data Extraction and Quality Assessment

A data extraction form was created to identify the study population (age, sex, and targeted health condition), study characteristics (country where the study took place, recruitment method, language, sample size), external reference, and psychometric parameters (validity, reliability, and responsiveness). Any disagreements on the eligibility of a study were resolved by consensus.

The quality of the articles chosen for inclusion was determined by using a 13-item critical appraisal tool developed specifically to assess psychometric properties for items used in clinical practice.[31] The items are listed in Appendix 1. Of the 13 items, 4 were for articles assessing reliability, 4 were for validity studies, and the remaining 5 items were for both.[31] Two authors (A.R. and A.K.) independently screened each article with the critical appraisal tool. With each of the 13 items, the critical appraisal tool gives a justification as to why the criterion should be evaluated in articles and a scoring rubric to help decide whether the article met the criteria or not. The quality assessments for each article were compared and discrepancies resolved. Authors answered either yes/no/not applicable for each of the 13 items and scores $\geq$80% for the article was considered good quality (Table 1).

Data on the psychometric properties of validity, including accuracy parameters, reliability, and responsiveness to change were extracted from the articles included in the review.

Psychometric Properties Evaluated

Validity refers to the extent to which an item is measuring the construct it claims to measure.[52] The item can be compared to a diagnosis, or “gold standard,”[53] or an already validated and reliable measure of the construct.[53] Validity is assessed by parameters such as the area under the receiver operating characteristics (ROC) curve, and correlation coefficients, such as Pearson r and Spearman ρ, are the parameters used to quantify validity.[54] The area under the ROC is calculated from the plot of the sensitivity and 1-specificity at each cut-point of the measure.[53] For these, the parameters closest to 1.0 have the highest validity.

Test-retest reliability assesses the stability of patients’ responses over time, given that the patients have not changed.[52,54,55] Three parameters were commonly reported: the intraclass correlation coefficient (ICC) for continuous variables; correlation coefficients, mainly Pearson r; and Cohen κ for dichotomous or ordinal variables, as well as weighted κ for ordinal variables.[53]

Responsiveness, which can be considered an aspect of validity,[53] detects change over time.[52] If patients have changed on the construct being measured, their scores should reflect this change accordingly. Responsiveness of an item is determined by how well it captures this change in status.[53] Responsiveness is commonly reported using the effect size (ES). For all of the parameters mentioned, the closer the value was to 1.0, the more acceptable the psychometric property was for the single item.

Psychometric Property Cut-Offs for Interpretation

To interpret the measurement properties extracted from the articles, standardized criteria was used for the values of validity, reliability, and area under the ROC curve.

Several authors have strength criteria for validity. However, it is all dependent on context and here we chose to use Cohen κ, where a correlation of 0.20 was small, 0.50 was moderate, and 0.80 was strong.[56] For interpreting the area under the ROC curve, which ranges from 0 to 1.0 (perfect prediction), 0.5 was poor (equivalent to predicting by flipping a coin), >0.7 indicated acceptable predictions, and >0.8 excellent predictions.[57] For κ, (weighted or not)

TABLE 1. Quality Assessment of Articles

	References	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13
Stress	Bulli et al[32]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	Goebel and Mehdorn[33]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y
	Gunnarsdottir et al[34]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	Hegel et al[35]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	Jacobsen et al[36]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y
	Keir et al[37]	Y	N/A	N	N/A	N/A	N/A	Y	N/A	Y	Y	N	Y	Y
	Thekkumpurath et al[38]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
Depression	Akechi et al[39]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	Chochinov et al[40]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y
	Kawase et al[41]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	N	Y
	Ayalon et al[42]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y
	Jesse and Graham[43]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	Watkins et al[44]	Y	N/A	Y	N/A	N/A	N/A	N	N/A	Y	Y	Y	Y	Y
Anxiety	Chlan[14]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	De Jong et al[44]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	Elliot[46]	N	N/A	Y	N	N/A	N/A	Y	N	Y	Y	Y	N	Y
	McKinley et al[16]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	McKinley and Madronio[17]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
Fatigue	Schwartz et al[47]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
	van Hooff et al[48]	Y	N/A	Y	N/A	N/A	N/A	Y	N/A	Y	Y	Y	N	Y
Pain	ten Klooster et al[49]	Y	N/A	N	N/A	N	N/A	Y	Y	Y	Y	Y	Y	Y
GHP	DeSalvo et al[15]	Y	N/A	Y	N/A	N	N/A	Y	Y	Y	Y	Y	Y	Y
	Rohrer et al[50]	Y	N/A	N	N/A	N/A	N/A	Y	N/A	Y	Y	Y	Y	Y
Sleep	Cappelleri et al[51]	Y	N/A	Y	N/A	N	N/A	Y	N	Y	Y	Y	Y	Y

GHP = general health perception, N/A = not applicable.

values of 0.75 or greater were considered excellent agreement.[58] As for responsiveness, an ES of 0.2 was considered weak, 0.5 moderate, and 0.8 strong.[56]

RESULTS

Figure 1 presents the flowchart of the literature search. The 7 constructs were searched and a total of 773 articles were found. After duplicates were removed (n = 360), a total of 413 abstracts were identified through the different databases. Of these, 318 abstracts were excluded because they were irrelevant, did not include a single-item PRO, or were unpublished data. Remaining articles were then evaluated for eligibility. Seventy-one articles were excluded upon full text review because they did not meet the single-item inclusion criteria (ie, not relating to the construct, not assessing symptom severity, no simple response options), whereas other articles were excluded because of no external reference standard. A total of 24 eligible articles were reviewed systematically. The literature on the PROMIS global items was not included as they had been validated against another single-item PRO question, namely, the classifiers of the EQ-5D.[29] Also, the VAMS items were not included because they were validated as a total score and not as standalone items.[28]

The 24 studies[14–17,50–51] included in our review were from countries all over the globe, including but not limited to Japan, the USA, Israel, several European countries, and Canada. The main study populations were patients with cancer,[32–37,39–41,47] stroke,[44] and myocardial infarction,[45,46] and seniors in primary care settings.[42] The sample sizes ranged from as low[50] as 34 to as high as 1493.[51]

General Health Perception

One item for general health perception was found from 2 articles. "How is your health in general?"[15,50] showed weak to moderate strength correlations ($r = 0.37$–0.66) when compared with the Short Form-12, Thomson Reuters, and a test–retest reliability of ICC = 0.69 (Table 2). The articles[59] had good methodological quality.

Stress

Stress was evaluated in 7 articles using the distress thermometer (DT), a visual analog scale (VAS) (Table 3).[32–37] This item was compared with several self-reported questionnaires, for example, the Patient Health Questionnaire. With DT cut-off scores ranging from 3 to 8, the area under the ROC ranged from fair to good (AUC = 0.73–0.88).[32–37] The item also showed a range of correlational strengths, although most were weak to moderate ($r = 0.23$–0.60).[32–37] All but 1 of the articles were of good quality.[37]

Pain

Two pain single items from 1 article were included in the review (Table 4). The first was a graphic rating scale

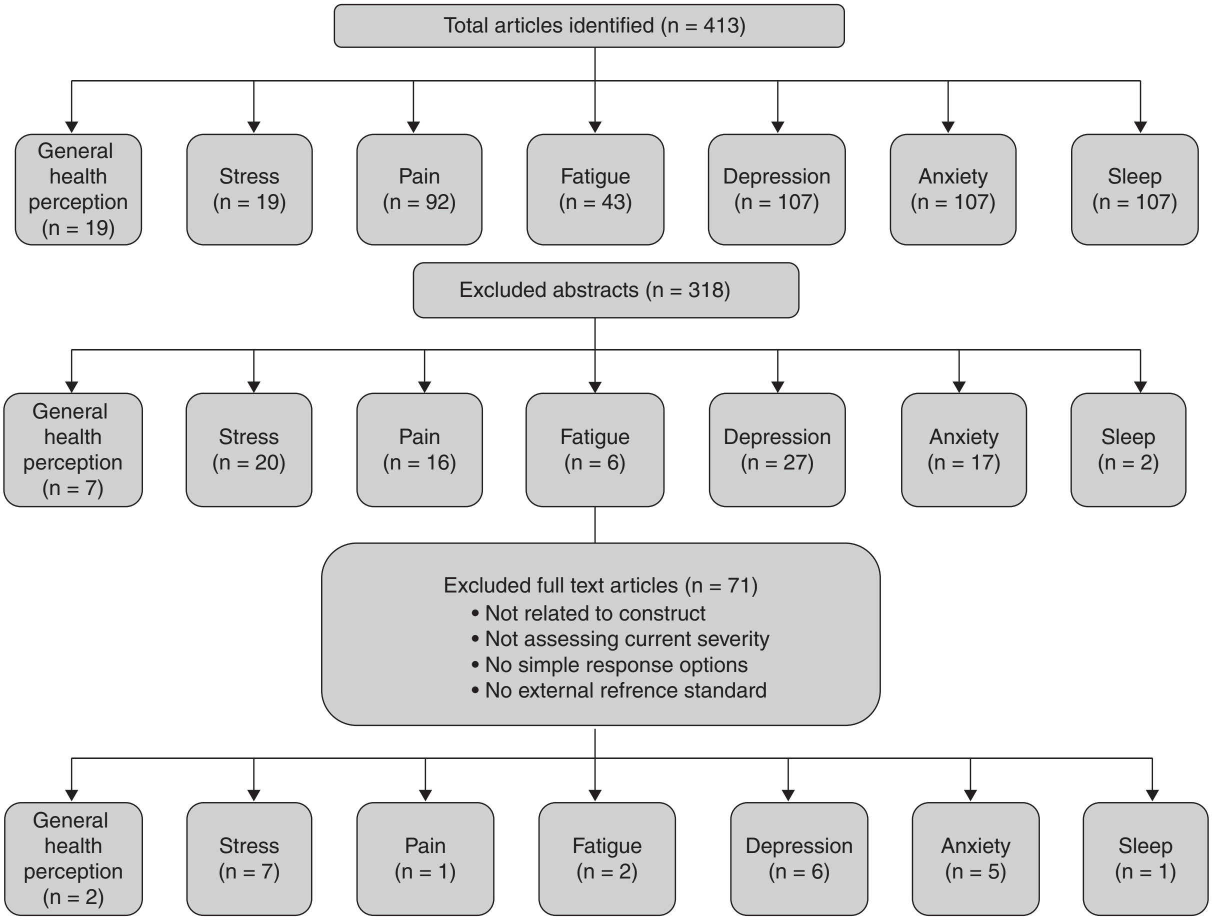

Total articles identified (n = 413)
General health perception (n = 19)
Stress (n = 19)
Pain (n = 92)
Fatigue (n = 43)
Depression (n = 107)
Anxiety (n = 107)
Sleep (n = 107)
Excluded abstracts (n = 318)
General health perception (n = 7)
Stress (n = 20)
Pain (n = 16)
Fatigue (n = 6)
Depression (n = 27)
Anxiety (n = 17)
Sleep (n = 2)
Excluded full text articles (n = 71)
• Not related to construct
• Not assessing current severity
• No simple response options
• No external refrence standard
General health perception (n = 2)
Stress (n = 7)
Pain (n = 1)
Fatigue (n = 2)
Depression (n = 6)
Anxiety (n = 5)
Sleep (n = 1)

FIGURE 1. Flowchart of the literature search.

TABLE 2. General Health Perception*

Author (Quality Assessment Value)	Sample Size	Study Sample	Validity: Correlation (Pearson *r*)	Reliability (ICC)
DeSalvo et al[15] (9/10)	$N_{reliability} = 75$ $N_{validity} = 104$	Veterans	0.65–0.66	0.69
Rohrer et al[50] (7/8)	34	People looking to lose weight	0.37–0.39	—
Total	138		0.37–0.66	0.69

ICC = intraclass correlation coefficient.
*How is your health in general? Excellent, very good, good, fair, poor.

(GRS) of pain severity, and the second was a verbal rating scale (VRS) where the response option is categorical (none to severe).[49] Both the GRS and VRS showed acceptable test–retest reliability (ICC = 0.85, $\kappa = 0.76$, respectively) and ranged from weak to moderate correlations ($r = 0.24–0.65$) when compared with other pain measures.[49] The article had good methodological quality.

Fatigue

Two items from 2 articles showed suitable clinical application (Table 5). Both of the items used the VAS metric, but different pegs: 0 (no fatigue) to 10 (greatest possible fatigue)[47] and 1 (not at all fatigued) to 10 (extremely fatigued).[48] "What is your current level of fatigue today?"[47] had a large effect size (0.78) and "How fatigued do you currently feel?"[48] showed weak to strong correlations ($r = 0.16$ 0.80) when compared with self-reported measures, such as the daily fatigue item in the Profile of Mood States. One of the articles demonstrated good discriminant validity ($r = -0.02$ to 0.10)[48] when compared against constructs unrelated to fatigue. Both the articles showed good methodological quality.

Depression

Three variations of single items from 6 articles were included for depression single item: "Are you depressed?"[39–41]; "Do you think you suffer from depression?"[42]; and "Are you often sad or depressed?"[43,44] (Table 6). The response options were binary (yes/no). The items, compared with structured clinical interviews and validated depression questionnaires, showed good accuracy with the area under the ROC curve of 0.85 and large sensitivities and specificities (sensitivity = 0.42–1.0, specificity = 0.60–1.0).[39–44] Of the 6 depression articles, 5 had good methodological quality, whereas 1 of them had moderate methodological quality.[41]

Anxiety

Two items from 5 articles were included to assess anxiety: a VAS[14,45,46] and a face anxiety scale (FAS) (Table 7).[16,17] When compared with anxiety questionnaires, the VAS showed a range of correlations from weak to moderate strength ($r = 0.28–0.70$).[14,45,46] The FAS showed moderate strength correlations when compared with interviews and self-reported anxiety scales ($r = 0.64–0.70$).[16,17] Four of the 5 articles were of good quality,[14,16,17,45] whereas 1 showed moderate methodological quality.[46]

Sleep

The Sleep quality scale was the single item included[51] (Table 8). This item showed acceptable test–retest reliability (ICC = 0.9); however, Pearson correlations were weak when compared with different subscales of the Medical Outcomes

TABLE 3. Stress*

			Validity					
Author (Quality Assessment Value)	Sample Size	Study Sample	AUC	Sensitivity	Specificity	PPV	NPV	Correlation (Pearson *r*)
Bulli et al[32] (8/8)	290	Cancer patients	0.84	0.73	0.82	0.69	0.85	—
Goebel and Mehdorn[33] (7/8)	150	Intracranial tumor patients	0.83–0.88	0.82–1.0	—	—	—	—
Gunnarsdottir et al[34] (8/8)	149	Cancer patients	0.74–0.75	0.72–0.77	0.61–0.69	—	—	0.45–0.57
Hegel et al[35] (8/8)	321	Cancer patients	0.86	0.81	0.85	—	—	—
Jacobsen et al[36] (7/8)	380	Cancer patients	0.78–0.80	0.70–0.77	0.68–0.70	—	—	—
Keir et al[37] (6/8)	75	Brain tumor patients	—	—	—	—	—	0.23–0.44
Thekkumpurath et al[38] (8/8)	150	Palliative care patients	0.729	0.77	0.59	0.49	0.94	—
Total	1515		0.74–0.80	0.70–0.77	0.61–0.70	0.49–0.69	0.85–0.94	0.23–0.57

AUC = area under the ROC curve, DT = distress thermometer, NPV = negative predictive value, PPV = positive predictive value, ROC = receiver operating characteristic.
*DT: Circle the number on the thermometer that best represents how much distress you have been experiencing during the last week including today, from 0 (no distress) to 10 (extreme distress).

TABLE 4. Pain

ten Klooster et al[49] (8/10)*	72	Rheumatology patients	0.28–0.65	ICC = 0.85
ten Klooster et al[49] (8/10)[†]	72	Rheumatology patients	0.24–0.59	κ = 0.76
Author (Quality Assessment Value)	Sample Size	Study Sample	Validity: Correlation (Spearman ρ)	Reliability

GRS = graphic rating scale, ICC = intraclass correlation coefficient, VRS = verbal rating scale.
*GRS: Mark the line at a point that best represents the severity of your pain, where 0 is no pain and 100 is severe pain.
[†]VRS: Select the word that best describes your usual pain: 1, none; 2, very mild; 3, mild; 4, moderate; 5, severe.

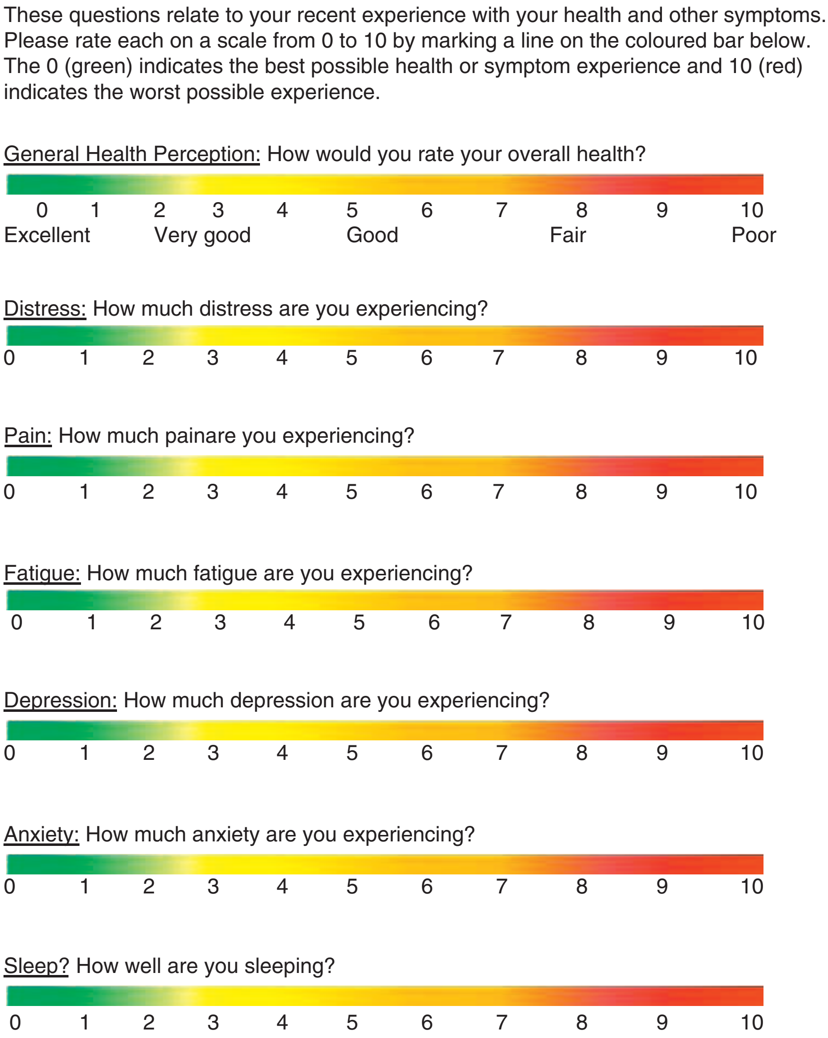
Study sleep scale domains ($r = 0.11–0.45$). The article had good methodological quality.

VAHS: How Are You Today?

Although there was a range of values for psychometric properties, for each domain assessed, there was at least 1 study showing strong relationship with the intended construct, all studies supporting a positive relationship (Figure 2). There is sufficient validity for these single items to warrant asking them in clinical practice. However, no one response set was tested. The most frequent option was the VAS. The VAS is widely used as a metric and has had a presence in the literature for almost a century.[60,61] By compiling the 7 single items identified in this literature review and using the VAS metric, we have created the visual analog health state (VAHS) form, which is presented in Figure 2.

DISCUSSION

The purpose of this review was to provide a clinically relevant bank of reliable and valid single-item outcome



These questions relate to your recent experience with your health and other symptoms.
Please rate each on a scale from 0 to 10 by marking a line on the coloured bar below.
The 0 (green) indicates the best possible health or symptom experience and 10 (red) indicates the worst possible experience.

General Health Perception: How would you rate your overall health?

0 1 2 3 4 5 6 7 8 9 10
Excellent Very good Good Fair Poor

Distress: How much distress are you experiencing?

0 1 2 3 4 5 6 7 8 9 10

Pain: How much painare you experiencing?

0 1 2 3 4 5 6 7 8 9 10

Fatigue: How much fatigue are you experiencing?

0 1 2 3 4 5 6 7 8 9 10

Depression: How much depression are you experiencing?

0 1 2 3 4 5 6 7 8 9 10

Anxiety: How much anxiety are you experiencing?

0 1 2 3 4 5 6 7 8 9 10

Sleep? How well are you sleeping?

0 1 2 3 4 5 6 7 8 9 10

FIGURE 2. How are you today? Visual analog health states.

TABLE 5. Fatigue*

Author (Quality Assessment Value)	Sample Size	Study Sample	Validity: Correlation (Pearson r)	Responsiveness (ES)
Schwartz et al[47] (8/8)	103	Cancer patients	—	0.78
van Hooff et al[48] (7/8)	130	Academic staff members of a Dutch university	Convergent: 0.16–0.80 Discriminant: 0.02–0.10	—
Total	233		Convergent: 0.16–0.80 Discriminant: 0.02–0.10	0.78

ES = effect size.

*Two variations of the same item: What is your level of fatigue today? Where 0 is no fatigue and 10 is the greatest possible fatigue. How fatigued do you currently feel? From 1, not at all fatigued, to 10, extremely fatigued.

TABLE 6. Depression*

			Validity				
Author (Quality Assessment Value)	Sample Size	Study Sample	AUC	Sensitivity	Specificity	PPV	NPV
Akechi et al[39] (8/8)	209	Terminally ill cancer patients	0.85	0.79	0.92	0.41	0.98
Ayalon et al[42] (7/8)	153	Primary care patients	—	0.83	0.83	0.17	0.99
Chochinov et al[40] (7/8)	197	Terminally ill cancer patients	—	1.0	1.0	1.0	1.0
Jesse and Graham[43] (8/8)	130	Pregnant low-income women	—	0.80	0.60	0.43	0.89
Kawase et al[41] (6/8)	282	Cancer patients receiving radiotherapy	—	0.42	0.86	0.22	0.94
Watkins et al[44] (7/8)	122	Stroke patients	—	0.86–0.95	0.84–0.89	0.86–0.93	0.84–0.91
Total	1093		0.85	0.42–1.0	0.60–1.0	0.22–1.0	0.84–1.0

AUC = area under the ROC curve, NPV = negative predictive value, PPV = positive predictive value.

*Three variations of the same item: Are you depressed? Do you think you suffer from depression? Are you often sad or depressed?

TABLE 7. Anxiety

Author (Quality Assessment Value)	Sample Size	Study Sample	Validity: Correlation (Pearson r)	Reliability: Correlation (Pearson r)*
Chlan[14] (8/8)	200	Critically ill patients receiving mechanical ventilation	0.50	—
De Jong et al[45] (8/8)	243	Acute myocardial infarction patients	0.45–0.52	—
Elliot[46] (6/10)	56	Acute myocardial infarction or unstable angina pectoris patients	0.28–0.70	0.61
Total	499		0.28–0.70	0.61
				Validity: Correlation (Spearman ρ)[†]
McKinley et al[16] (8/8)	106	Critically ill patients that are mechanically ventilated	0.64	—
McKinley and Madronio[17] (8/8)	100	Critically ill patients that are mechanically ventilated	—	0.70
Total	206		0.64	0.70

FAS = face anxiety scale, VAS = visual analog scale.

*We are interested in how anxious you feel now. On a VAS from 0 to 10, where 0 is no anxiety and 10 is the most anxiety you have ever experienced, please rate your current anxiety level.

[†]FAS: These faces are showing different levels of anxiety. This face shows no anxiety at all, this face shows a little bit more, a bit more, right up to extreme anxiety. Have a look at the faces and choose the one that shows how much anxiety you are feeling at the moment.

TABLE 8. Sleep*

Cappelleri et al[51] (8/10)	1493	Fibromyalgia patients	0.11–0.45	0.9
Author (Quality Assessment Value)	Sample Size	Study Sample	Validity: Correlation (Pearson r)	Reliability (ICC)

ICC = intraclass correlation coefficient.
*Sleep quality scale: Rate the quality of your sleep over the past 24 hours, from 0 (best possible sleep) to 10 (worst possible sleep).

measures that could be used to efficiently and effectively structure the patient/clinician dialogue around patient-centered outcomes. To achieve this, the psychometric properties of the single items were reviewed and there was support for their validity as appropriate questions to ask in a clinical setting.

There are several advantages associated with the use of single items over multi-item questionnaires. They are simple and easy to implement, more time efficient,[18] inexpensive,[20] and can be more appropriate in certain patient populations.[19] Conversely, questionnaires that have multiple items can be time-consuming for patients as well as require significant concentration and attention. Regardless of the disease population, single items can act in the best interest of patients in that they respect patients' time and can access valuable data without a cognitively demanding multi-item questionnaire.[19] These items can be used to monitor changes in health states through the review of the single-item score over time.[18,20] Additionally, the clinical utility of single items serves to maximize the time spent on interviews or physical evaluations and complement the overall assessment without interfering with the lengthy routine of clinical process.[18] Furthermore, there are specific patient populations in which single-item measures are the only appropriate option, such as critically ill patients, where it can be difficult to evaluate PROs with interviews or lengthy questionnaires.[19]

The literature supports the usefulness of single items. Our search revealed several examples of valid single-item variations for use in clinical settings. Specifically, the single item for general health perception was significantly correlated with tumor necrosis factor-α, interleukin-6, and C-reactive protein levels for all age groups, where worse health correlated with higher levels of the inflammatory cytokines[62–64] as well as other circulating biomarkers.[65] A single question added into a clinical encounter can provide both a self-reported psychological and biological marker of a patient's health status.[62] If the patient reports poor self-rated health, the physician can use this as an indicator of a potential underlying illness that may otherwise go unnoticed.[62]

Asking relevant questions is part of etiquette-based medicine.[66] Eliciting patients' concerns can help emphasize feelings of respect and present the physician as more courteous and humane.[66,67] Some of these behaviors are as simple as introducing yourself and shaking the patient's hand when entering the room.[66] This greatly increases patient satisfaction, reduces anxiety, and increases adherence to medication or treatment[68,69]; however, these behaviors are not always performed.[70] An important component of etiquette-based medicine is the use of open questions rather than yes/no questions to elicit feelings.[66] Optimal behavior would also include paying attention, reporting the answers, and using them for future comparison. We propose that by asking patients about the 7 domains mentioned, patient-centered care could be improved. The VAHS gives a quick profile of a patient's PRO health state, which provides useful information about the patient at the present time and can be used as a reference point for monitoring during the course of treatment.

A limitation in our review is that the identification of articles validating single items was not uncomplicated because there is no consistent way in which the information is indexed. As a result, articles were likely missed. Enhancing the article selection with known literature was an attempt to fill some gaps. Hence, we looked at the literature on VAMS[28] and PROMIS,[29] but it was not retained for the analysis because the method of cross-validation did not meet our criteria. These 2 sources, however, support the value in using single items in general.

Furthermore, we specifically targeted articles on the psychometric properties of these measures. There may have been additional sources that provided information on validity and reliability but not as the primary objective of the study. As well, there are limitations when using only 1 item to assess complicated outcomes. Multi-item questionnaires have more consistency, are less susceptible to bias,[18] and may provide more information than just 1 self-reported item. There are advantages and disadvantages to both single and multi-item PRO questionnaires, and there is evidence for the usefulness of both. It is important to realize the contextual and situational uses for each one.

We chose the VAS as the response set for the VAHS. Psychometric properties of the VAS have been frequently studied for mood and pain.[69] Validity shows correlation strengths ranging from moderate to strong for pain,[72–74] and weak to moderate for mood.[75] The VAS demonstrates an ability to discern small decreases in pain in a clinical setting,[74,76] as well as discriminate between different levels of pain intensity.[76] Although quick and easily administered, there are limitations to its use. The VAS metric can vary with experience,[71] for example, a maximal value of pain for an individual can change if, between the 2 VAS time-point measurements, the patient has a painful experience.[71] Despite these limitations, the VAS has demonstrated strong evidence of validity, simplicity, and clinical usefulness in a prospective manner. Alternatively, a Likert scale with verbal responses (not at all, a little bit, somewhat, quite a bit, very much) is a common method for scoring single items,[29] but the interpretation of the qualifiers may not be the same across people, health conditions, and languages.

In conclusion, we recommend that clinicians use these validated items in their clinical practice to enhance patient-centered care and permit tracking of a patient's progress over time. Future research should focus on evaluating the impact of using such a reporting system on patient and clinician satisfaction, as well as adherence to treatment.

APPENDIX 1

Quality Assessment of Articles

1. Item 1: If human subjects were used, did the authors give a detailed description of the sample of subjects used to perform the (index) test?

2. Item 2: Did the authors clarify the qualification, or competence of the rater(s) who performed the (index) test?
3. Item 3: Was the reference standard explained?
4. Item 4: If interrater reliability was tested, were raters blinded to the findings of other raters?
5. Item 5: If intrarater reliability was tested, were raters blinded to their own prior findings of the test under evaluation?
6. Item 6: Was the order of examination varied?
7. Item 7: If human participants were used, was the time period between the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the 2 tests?
8. Item 8: Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time interval between repeated measures?
9. Item 9: Was the reference standard independent to the index test?
10. Item 10: Was the execution of the (index) test described in sufficient detail to permit replication of the test?
11. Item 11: Was the execution of the reference standard described in sufficient detail to permit its replication?
12. Item 12: Were withdrawals from the study explained?
13. Item 13: Were the statistical methods appropriate for the purpose of the study?

REFERENCES

1. Patient-Centered Outcomes Research Institute. *Patient-Centered Outcomes Research*. Washington, DC: Patient-Centered Outcomes Research Institute; 2013.
2. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *J Eval Clin Pract*. 1999;5:401–416.
3. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res*. 1995;4:293–307.
4. Velikova G, Wright P. Individual patient monitoring. *Assessing Quality of Life in Clinical Trials*. 2nd ed. New York: Oxford University Press; 2005:291–306.
5. Bennett AV, Jensen RE, Basch E. Electronic patient-reported outcome systems in oncology clinical practice. *CA Cancer J Clin*. 2012;62:337–347.
6. Snyder CF, Herman JM, White SM,et al. When using patient-reported outcomes in clinical practice, the measure matters: a randomized controlled trial. *J Oncol Pract*. 2014;10:e299–e306.
7. Rose M, Bezjak A. Logistics of collecting patient-reported outcomes (PROs) in clinical practice: an overview and practical examples. *Qual Life Res*. 2009;18:125–136.
8. Paterson C. Measuring outcomes in primary care: a patient generated measure, MYMOP, compared with the SF-36 health survey. *BMJ*. 1996;312:1016–1020.
9. Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Soc Sci Med*. 2005;60:833–843.
10. Higginson IJ, Carr AJ. Measuring quality of life: using quality of life measures in the clinical setting. *BMJ*. 2001;322:1297–1300.
11. Joyce CRB, McGee HM, O'Boyle CA. *Individual Quality of Life: Approaches to Conceptualisation and Assessment*. Amsterdam, The Netherlands: Taylor & Francis; 1999.
12. Fayers P, Hays R. Evaluating mult-item scales. *Assessing Quality of Life in Clinical Trials*. 2nd ed. New York: Oxford University Press; 2005.
13. Greenhalgh J. The applications of PROs in clinical practice: what are they, do they work, and why? *Qual Life Res*. 2009;18:115–123.
14. Chlan LL. Relationship between two anxiety instruments in patients receiving mechanical ventilatory support. *J Adv Nurs*. 2004;48:493–499.
15. DeSalvo KB, Fisher WP, Tran K, et al. Assessing measurement properties of two single-item general health measures. *Qual Life Res*. 2006;15:191–201.
16. McKinley S, Stein-Parbury J, Chehelnabi A, et al. Assessment of anxiety in intensive care patients by using the Faces Anxiety Scale. *Am J Crit Care*. 2004;13:146–152.
17. McKinley S, Madronio C. Validity of the Faces Anxiety Scale for the assessment of state anxiety in intensive care patients not receiving mechanical ventilation. *J Psychosom Res*. 2008;64:503–507.
18. Bowling A. Just one question: if one question works, why ask several? *J Epidemiol Community Health*. 2005;59:342–345.
19. McKinley S, Coote K, Stein-Parbury J. Development and testing of a Faces Scale for the assessment of anxiety in critically ill patients. *J Adv Nurs*. 2003;41:73–79.
20. Surti B, Spiegel B, Ippoliti A, et al. Assessing health status in inflammatory bowel disease using a novel single-item numeric rating scale. *Dig Dis Sci*. 2013;58:1313–1321.
21. United States Food and Drug Administration. Guidance for industry on patient-reported outcome measures—use in medical product development to support labeling claims. *Federal Registry*. 2009;74:65132–65133.
22. Chalut DS, Ducharme FM, Davis GM. The Preschool Respiratory Assessment Measure (PRAM): a responsive index of acute asthma severity. *J Pediatr*. 2000;137:762–768.
23. Mattis S. *Dementia Rating Scale*. Odessa, FL: Psychological Assessment Resources; 1988.
24. Cameron JI, Cheung AM, Streiner DL, et al. Factor structure and reliability of the brain impairment behavior scale. *J Neurosci Nurs*. 2008;40:40–47.
25. Bogousslavsky J. William Feinberg lecture 2002: emotions, mood, and behavior after stroke. *Stroke*. 2003;34:1046–1050.
26. Guyatt GH, Sullivan MJ, Thompson PJ, et al. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. *Can Med Assoc J*. 1985;132:919–923.
27. Mahoney FI, Barthel DW. Functional evaluation: The Barthel lndex. *Md St Med J*. 1965;14:61–65.
28. Cella DF, Perry SW. Reliability and concurrent validity of three visual-analogue mood scales. *Psychol Rep*. 1986;59:827–833.
29. Revicki DA, Kawata AK, Harnam N, et al. Predicting EuroQol (EQ-5D) scores from the patient-reported outcomes measurement information system (PROMIS) global items and domain item banks in a United States sample. *Qual Life Res*. 2009;18:783–791.
30. Mayo NE, Hum S, Kuspinar A. Methods and measures: what's new for MS? *Mult Scler*. 2013;19:709–713.
31. Brink Y, Louw QA. Clinical instruments: reliability and validity critical appraisal. *J Eval Clin Pract*. 2012;18:1126–1132.
32. Bulli F, Miccinesi G, Maruelli A, et al. The measure of psychological distress in cancer patients: the use of Distress Thermometer in the Oncological Rehabilitation Center of Florence. *Support Care Cancer*. 2009;17:771–779.
33. Goebel S, Mehdorn HM. Measurement of psychological distress in patients with intracranial tumours: the NCCN distress thermometer. *J Neurooncol*. 2011;104:357–364.
34. Gunnarsdottir S, Thorvaldsdottir GH, Fridriksdottir N, et al. The psychometric properties of the Icelandic version of the distress thermometer and problem list. *Psychooncology*. 2012;21:730–736.
35. Hegel MT, Collins ED, Kearing S, et al. Sensitivity and specificity of the Distress Thermometer for depression in newly diagnosed breast cancer patients. *Psychooncology*. 2008;17:556–560.

36. Jacobsen PB, Donovan KA, Trask PC, et al. Screening for psychologic distress in ambulatory cancer patients. *Cancer*. 2005;103:1494–1502.

37. Keir ST, Calhoun-Eagan RD, Swartz JJ, et al. Screening for distress in patients with brain cancer using the NCCN's rapid screening measure. *Psychooncology*. 2008;17:621–625.

38. Thekkumpurath P, Venkateswaran C, Kumar M, et al. Screening for psychological distress in palliative care: performance of touch screen questionnaires compared with semistructured psychiatric interview. *J Pain Symptom Manage*. 2009;38:597–605.

39. Akechi T, Okuyama T, Sugawara Y, et al. Screening for depression in terminally ill cancer patients in Japan. *J Pain Symptom Manage*. 2006;31:5–12.

40. Chochinov HM, Wilson KG, Enns M, et al. "Are you depressed?" Screening for depression in the terminally ill. *Am J Psychiatry*. 1997;154:674–676.

41. Kawase E, Karasawa K, Shimotsu S, et al. Evaluation of a one-question interview for depression in a radiation oncology department in Japan. *Gen Hosp Psychiatry*. 2006;28:321–322.

42. Ayalon L, Goldfracht M, Bech P. 'Do you think you suffer from depression?' Reevaluating the use of a single item question for the screening of depression in older primary care patients. *Int J Geriatr Psychiatry*. 2010;25:497–502.

43. Jesse DE, Graham M. Are you often sad and depressed? Brief measures to identify women at risk for depression in pregnancy. *MCN Am J Matern Child Nurs*. 2005;30:40–45.

44. Watkins CL, Lightbody CE, Sutton CJ, et al. Evaluation of a single-item screening tool for depression after stroke: a cohort study. *Clin Rehabil*. 2007;21:846–852.

45. De Jong MM, An K, McKinley S, et al. Using a 0–10 scale for assessment of anxiety in patients with acute myocardial infarction. *Dimens Crit Care Nurs*. 2005;24:139–146.

46. Elliot D. Comparison of three instruments for measuring patient anxiety in a coronary care unit. *Intensive Crit Care Nurs*. 1993;9:195–200.

47. Schwartz AL, Meek PM, Nail LM, et al. Measurement of fatigue: determining minimally important clinical differences. *J Clin Epidemiol*. 2002;55:239–244.

48. van Hooff ML, Geurts SA, Kompier MA, et al. "How fatigued do you currently feel?" Convergent and discriminant validity of a single-item fatigue measure. *J Occup Health*. 2007;49:224–234.

49. ten Klooster PM, Vlaar AP, Taal E, et al. The validity and reliability of the graphic rating scale and verbal rating scale for measuring pain across cultures: a study in Egyptian and Dutch women with rheumatoid arthritis. *Clin J Pain*. 2006;22:827–830.

50. Rohrer JE, Herman DC, Merry SP, et al. Validity of overall self-rated health as an outcome measure in small samples: a pilot study involving a case series. *J Eval Clin Pract*. 2009;15:366–369.

51. Cappelleri JC, Bushmakin AG, McDermott AM, et al. Psychometric properties of a single-item scale to assess sleep quality among individuals with fibromyalgia. *Health Qual Life Outcomes*. 2009;7:54.

52. Mokkink LB, Terwee CB, Gibbons E, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Med Res Methodol*. 2010;10:82.

53. de Vet H, Terwee C, Mokkink L, et al. *Measurement in Medicine*. Cambridge: Cambridge University Press; 2011.

54. Finch E, Brooks D, Stratford P, et al. *Physical Rehabilitation Outcome Measures: A Guide to Enhanced Clinical Decision Making*. 2nd ed. Toronto, ON: Canadian Physiotherapy Association; 2002.

55. Feldt L, Brennan R. Reliability. In: Linn R, ed. *Educational Measurement*. Pheonix: Oryx Press; 1993;105–146.

56. Cohen J. A power primer. *Psychol Bull*. 1992;112:155–159.

57. Schneeweiss S, Seeger JD, Maclure M, et al. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am J Epidemiol*. 2001;154:854–864.

58. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*. 1977;33:363–374.

59. Ware Jr. JE, Kosinski M, Keller SD. A 12-item short-form health survey. Construction of scales and preliminary tests of reliability and validity. *Med Care.* 1996;34:220–233.

60. Freyd M. The Graphic Rating Scale. *J Educ Psychol*. 1923;14: 83–102.

61. Hayes M, Patterson D. Experimental development of the graphic rating method. *Psychol Bull*. 1921;18:98–113.

62. Unden AL, Andreasson A, Elofsson S, et al. Inflammatory cytokines, behaviour and age as determinants of self-rated health in women. *Clin Sci (Lond)*. 2007;112:363–373.

63. Christian LM, Glaser R, Porter K, et al. Poorer self-rated health is associated with elevated inflammatory markers among older adults. *Psychoneuroendocrinology*. 2011;36:1495–1504.

64. Tanno K, Ohsawa M, Onoda T, et al. Poor self-rated health is significantly associated with elevated C-reactive protein levels in women, but not in men, in the Japanese general population. *J Psychosom Res*. 2012;73:225–231.

65. Jylha M, Volpato S, Guralnik JM. Self-rated health showed a graded association with frequently used biomarkers in a large population sample. *J Clin Epidemiol*. 2006;59:465–471.

66. Kahn MW. Etiquette-based medicine. *New Engl J Med*. 2008;358:1988–1989.

67. Block L, Hutzler L, Habicht R, et al. Do internal medicine interns practice etiquette-based communication? A critical look at the inpatient encounter. *J Hosp Med*. 2013;8:631–634.

68. Fogarty LA, Curbow BA, Wingard JR, et al. Can 40 seconds of compassion reduce patient anxiety? *J Clin Oncol*. 1999;17: 371–379.

69. Griffith CH III, Wilson JF, Langer S, et al. House staff nonverbal communication skills and standardized patient satisfaction. *J Gen Intern Med*. 2003;18:170–174.

70. Tackett S, Tad-y D, Rios R, et al. Appraising the practice of etiquette-based medicine in the inpatient setting. *J Gen Intern Med*. 2013;7:908–913.

71. Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. *Res Nurs Health*. 1990;13:227–236.

72. Ahles TA, Ruckdeschel JC, Blanchard EB. Cancer-related pain—II. Assessment with visual analogue scales. *J Psychosom Res*. 1984;28:121–124.

73. Downie WW, Leatham PA, Rhind VM, et al. Studies with pain rating scales. *Ann Rheum Dis*. 1978;37:378–381.

74. Seymour RA. The use of pain scales in assessing the efficacy of analgesics in post-operative dental pain. *Eur J Clin Pharmacol*. 1982;23:441–444.

75. Folstein MF, Luria R. Reliability, validity, and clinical application of the visual analogue mood scale. *Psychol Med*. 1973;3:479–486.

76. Price DD, McGrath PA, Rafii A, et al. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*. 1983;17:45–56.