Contents lists available at ScienceDirect

Current Therapeutic Research

Non–Small Cell Lung Cancer Symptom Assessment Questionnaire: Psychometric Performance and Regulatory Qualification of a Novel Patient-Reported Symptom Measure

Donald M. Bushnell, MA [1,*], Thomas M. Atkinson, PhD [2], Kelly P. McCarrier, PhD [3], Astra M. Liepa, PharmD [4], Kendra P. DeBusk, PhD [5], Stephen Joel Coons, PhD [6], on behalf of the Patient-Reported Outcome Consortium's NSCLC Working Group

[1] *Patient-Centered Research, Evidera, Seattle, Washington*
[2] *Memorial Sloan Kettering Cancer Center, Patient-Reported Outcomes, Community-Engagement, and Language Core Facility, New York, New York*
[3] *Pharmerit International, Patient-Centered Outcomes, Bethesda, Maryland*
[4] *Eli Lilly and Company, Global Patient Outcomes and Real World Evidence, Indianapolis, Indiana*
[5] *Seattle Genetics, Global HEOR, Bothell, Washington*
[6] *Critical Path Institute, Patient-Reported Outcome Consortium, Tucson, Arizona*

ARTICLE INFO

ABSTRACT

Background: The Non–Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ) was developed to incorporate the patient's perspective into evaluation of clinical benefit in advanced non–small cell lung cancer trials and meet regulatory expectations for doing so. Qualitative evidence supported 7 items covering 5 symptom concepts.

Objective: This study evaluated measurement properties of the NSCLC-SAQ's items, overall scale, and total score.

Methods: In this observational cross-sectional study, a purposive sample of patients with clinician-diagnosed advanced non–small cell lung cancer, initiating or undergoing treatment, provided sociodemographic information and completed the NSCLC-SAQ, National Comprehensive Cancer Network/Functional Assessment of Cancer Therapy Lung Symptom Index (FLSI-17), and a Patient Global Impression of Severity item. Rasch analyses, factor analyses, and assessments of construct validity and reliability were completed.

Results: The 152 participants had a mean age of 64 years, 57% were women, and 87% where White. The majority were Stage IV (83%), 51% had an Eastern Cooperative Oncology Group performance status of 1 (32% performance status 0 and 17% performance status 2), and 33% were treatment naïve. Rasch analyses showed ordered thresholds for response options. Factor analyses demonstrated that items could be combined for a total score. Internal consistency (Cronbach $\alpha = 0.78$) and test–retest reliability (intraclass correlation coefficient = 0.87) were quite satisfactory. NSCLC-SAQ total score correlation was 0.83 with the National Comprehensive Cancer Network/Functional Assessment of Cancer Therapy Lung Symptom Index-17. The NSCLC-SAQ was able to differentiate between symptom severity levels and performance status (both *P* values < .001).

Conclusions: The NSCLC-SAQ generated highly reliable scores with substantial evidence of construct validity. The Food and Drug Administration's qualification supports the NSCLC-SAQ as a measure of symptoms in drug development. Further evaluation is needed on its longitudinal measurement properties and interepretation of meaningful within-patient score change. (Curr Ther Res Clin Exp. 2021; 82:XXX–XXX)

Introduction

Lung cancer is among the most common cancers in terms of incidence. It was estimated that more than 200,000 new cases of

* Address correspondence to: Donald M. Bushnell, MA, Evidera, 1417 4th Ave, Suite 501, Seattle, WA 98101.
E-mail address: don.bushnell@evidera.com (D.M. Bushnell).

lung cancer would be diagnosed in the United States during 2018. Lung cancer is also the leading cause of cancer-related mortality in the United States, with 150,000 deaths annually.[1] Although there are more than a dozen different kinds of lung cancer, the 2 main types are non–small cell lung cancer (NSCLC) and small cell lung cancer. Approximately 75% to 80% of lung cancers are of the non–small cell type.[2]

In the assessment of drug efficacy, cancer trials traditionally rely on primary end points that are biomarker-based (eg, radiographic assessment of tumor size to evaluate progression-free survival). However, this approach can miss important clinical benefit that can arise from the alleviation or avoidance of symptoms or functional limitations caused by the disease or its treatment. Recognizing this, US Food and Drug Administration (FDA) staff proposed that cancer clinical trials should include individual patient-reported measures of treatment-related symptomatic adverse events, physical function, and disease-related symptoms.[3] Hence, assessment of the core symptoms of NSCLC is a key component of a more comprehensive evaluation of clinical benefit in NSCLC treatment trials. Because it is only 1 component of a broader patient-reported outcome (PRO) measurement strategy, minimizing patient burden in terms of the number of items is critical. Although numerous patient-reported NSCLC measures exist (eg, Functional Assessment of Cancer Therapy-Lung,[4] European Organization for Research and Treatment of Cancer Quality of Life Lung-Specific Questionnaire,[5] Lung Cancer Symptom Scale,[6] and the M.D. Anderson Symptom Inventory Lung Cancer Module[7]), no single existing measure has been used consistently in clinical development programs. Furthermore, the existing measures are not exclusively NSCLC-related symptom measures because they include broad, noproximal concepts such as quality of life and/or selected treatment-related signs and symptoms that may become less relevant as treatment evolves. In addition, based on interpretation of the evidentiary expectations (eg, concept elicitation reports with transcripts, saturation grids, and item tracking matrices) at the time, it was believed that existing measures would fall short in satisfying the regulatory requirements of FDA's drug development tool qualification program.[8]

In response to this need for fit-for-purpose clinical outcome measures, the PRO Consortium's[9] NSCLC Working Group sponsored the development of a new PRO measure designed to assess the core disease-related symptoms that are important and relevant to persons with advanced NSCLC. This measure, named the Non-Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ), was developed with consideration of the recommendations and scientific best practices set forth in the FDA guidance for industry titled *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*, hereafter called the PRO Guidance,[10] and recent scientific literature for achieving content validity of PRO instruments.[11–14] In addition, in 2014, the FDA released the Guidance for Industry and FDA Staff: Qualification Process for Drug Development Tools (hereafter referred to as FDA Qualification Guidance).[8] Qualification, as defined by the FDA's Center for Drug Evaluation and Research, is a formal conclusion that the results obtained from the PRO measure within a stated context of use can be relied upon to have a specific interpretation and application in drug development and regulatory review.[8] Development and FDA qualification of the NSCLC-SAQ was the goal of the PRO Consortium's NSCLC Working Group.

To date, the development of the NSCLC-SAQ has included completion of systematic reviews of the NSCLC literature and existing PRO and clinician-reported outcome instruments; the formation of an expert panel of clinical and methodological experts to provide advice during multiple stages of the development process (eg, review of study protocols and results); the completion of qualitative concept elicitation interviews conducted with patients with

NSCLC to identify the symptom-related concepts that are most important and relevant to their experiences; a formal item-generation process in which evidence from the concept elicitation interviews, systematic literature reviews, and expert input was used to develop the content of the NSCLC-SAQ; qualitative cognitive interviews among participants with NSCLC to evaluate and refine the draft instrument; an electronic implementation assessment (by the ePRO Consortium's Instrument Migration Subcommittee) to assess the ability to implement the NSCLC-SAQ on all available and appropriate electronic data capture platforms; and a translatability assessment, conducted concurrently with the early cognitive interview process. Throughout the PRO measure development process, the NSCLC Working Group received iterative feedback from FDA to ensure that the NSCLC-SAQ would be fit-for-purpose for use in drug development. The extensive qualitative work demonstrating the relevance and importance of the NSCLC-SAQ's content to persons with advanced NSCLC and clinicians has been published.[15]

This process resulted in a developmental version of the NSCLC-SAQ. The primary aim of this study was to generate quantitative evidence regarding the measurement properties of the NSCLC-SAQ. This study was conducted from February 2015 through August 2016.

Methods

A purposive sample of participants with clinician-diagnosed NSCLC from US clinical sites participated in data collection. The target sample size was 150 participants. Eligibility criteria were designed to reflect common entry criteria for clinical trials in advanced NSCLC. Eligibility included at least 18 years of age; diagnosis of Stage IIIB or IV NSCLC; naïve to treatment (naïve to chemotherapy at the participant's current stage of NSCLC as of enrollment and had not received chemotherapy in the past 6 months for earlier stage disease) or had fully recovered from the adverse event or recovered to at least a Common Terminology Criteria for Adverse Events version 4.03 grade 1.[16] Furthermore, patients were required to be able to read, write, and speak English.

A centralized ethics committee process was used for the participating sites that were able to use a central application; clinics that required their own internal ethics committee approval were supported with the study documentation and monitored to assure appropriate approval was obtained before the initiation of any study activities at that site. No clinical interventions or investigational products were used in this study and no change in treatment was required to participate.

Staff at each clinic site used medical records to identify patients potentially eligible for participation. Recruitment quotas were established to oversample certain subgroups to enable considerations of factors that might influence symptom experience of patients in advanced NSCLC clinical trials. These quotas included at least 30% with Stage IV, at least 10% with Eastern Cooperative Oncology Group (ECOG) performance status 2 or higher, no fewer than 40% and no more than 60% of the study sample with a diagnosis of comorbid chronic obstructive pulmonary disease (COPD), and at least 30% being treatment naïve at the time of enrollment.

If eligible by record screen (eg, age, primary cancer diagnosis and stage, treatment history, prior adverse events, and current Common Terminology Criteria for Adverse Events grade information), potential participants were contacted by telephone or in-person and provided with information about the study using a standardized screening script. Those interested completed a short series of additional screening questions (eg, date of birth, substance abuse, previous study participation, language/reading problems, and availability) and, if eligible, were then scheduled for their enrollment visit (with an effort to schedule the enrollment
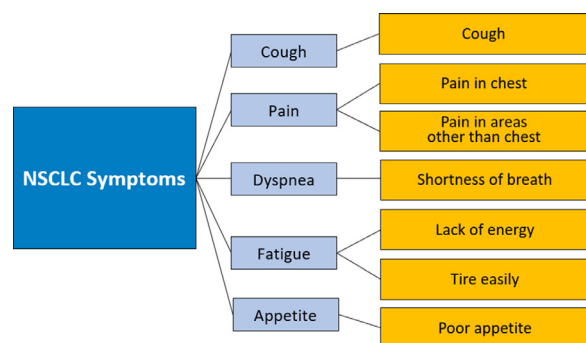
Figure 1. Conceptual framework for the Non–Small Cell Lung Cancer Symptom Assessment Questionnaire. NSCLC = non–small cell lung cancer.

visit to coincide with a regularly scheduled clinic visit for convenience).

On day 1, using a study tablet, each participant completed demographic items (marital status, education level, employment status, household income category, race, ethnicity, and self-reported general health) and the NSCLC-SAQ. Along with the NSCLC-SAQ, each participant completed the National Comprehensive Cancer Network/Functional Assessment of Cancer Therapy Lung Symptom Index-17 (FLSI-17)[17] and a single Patient Global Impression of Severity (PGIS) item. All participants were also expected to participate in a 1-week retest that included the NSCLC-SAQ and the PGIS item. The participants returned to the clinic for this visit to complete the measures on the study tablet. Responses to the retest were accepted within the window of 7 to 10 days, and participants were excluded from the retest analysis if outside this window.

Study outcome measures

The NSCLC-SAQ is a newly developed scale with seven items assessing 5 symptoms of NSCLC (ie, cough, pain, dyspnea, fatigue, and poor appetite) (see Figure 1).[15] The recall period is "over the last 7 days." Each item has a 5-point verbal rating scale from either 0 "No <symptom> at All" to 4 "Very severe <symptom>" or from 0 "Never" to 4 "Always," depending on the item's format (ie, intensity or frequency). Participants were allowed to skip NSCLC-SAQ items in the case that they actively indicated they did not wish to answer.

The FLSI-17[17] is a 17-item PRO instrument. All 17 items are used in computing a total score (range = 0–68 with 0 indicating a severely symptomatic patient), but the FLSI-17 is also evaluated in four areas associated with lung malignancies: Disease-Related Symptoms-Physical [DRS-P] (10 items, range = 0–40); Disease-Related Symptoms-Emotional (1 item, range = 0–4); Treatment Side Effects (3 items, range = 0–12); and Function/Well-Being (3 items, range = 0–12). The recall period is "the past 7 days." Each of the items uses a 5-point verbal rating scale in a tabular form, with options ranging from 0 ("Not at all") to 4 ("Very much"). The FLSI-17 was selected for use in this study because it has been used in NSCLC studies, provides coverage of symptom concepts similar to those included in the NSCLC-SAQ, and carries low respondent burden. It was hypothesized that the NSCLC-SAQ total score would be most highly associated with the FLSI-17 DRS-P score.

The PGIS is an assessment of lung cancer symptoms at the current time. It is a single item with the response options of: 0 = "Not severe," 1 = "Mildly severe," 2 = "Moderately severe," 3 = "Very severe," and 4 = "Extremely severe." The PGIS was used as the primary means for assessing change between the visits for the retest analysis.

Statistical analyses

Descriptive statistics (mean, standard deviation, median, and range for quantitative variables, and frequency and percentage for categorical variables) were calculated for demographic variables (eg, age, sex, race, ethnicity, marital status, education, employment status, and income) and for clinical variables (eg, current NSCLC stage, stage at initial NSCLC diagnosis, histology, years since NSCLC diagnosis, treatment status, ECOG performance status, clinical diagnosis of COPD, and smoking history) to describe the study sample.

The analyses used to evaluate the NSCLC-SAQ are described in detail in Table 1. The analyses used to evaluate NSCLC-SAQ item performance were in accordance with classical psychometric theory[18] and item response theory methods (ie, Rasch measurement theory [RMT] analysis).[19] Evaluation of the items in the NSCLC-SAQ was made using information from the following analyses: floor effect (where participants endorse the worst response option and can only improve) and ceiling effect (where participants endorse the best possible response option and can only get worse), item-to-item correlations, item-to-total correlations, factor analyses, reliability estimation, and item parameters from RMT analyses. All statistical tests used a significance level of 0.05 (2-sided) unless otherwise noted. Statistical tests involving multiple comparisons (eg, ANOVA models with multiple groups) included Scheffé post hoc tests, which adjust for multiple comparisons and reduce the possibility of Type I error. Statistics were conducted using SPSS version 18 (IBM-SPSS Inc, Armonk, NY),[20] RUMM2030 software (Rasch Unidimensional Measurement Model RUMM Laboratory Pty Ltd, Duncraig, Australia),[21] and IBM SPSS Amos (IBM SPSS Inc).[22,23]

Results

A total of 152 patients from 14 sites across 9 states (New York, Illinois, Alabama, Georgia, Ohio, Kentucky, Pennsylvania, Florida, and Louisiana) were enrolled into the study. Demographic and clinical characteristics are shown in Table 2. Participants were 57% women; 87% White; and, on average, aged 64 years. More than half (61%) were married or living as married, 84% had at least a high school education, 49% were retired, 18% were unable to work, and 54% had an annual household income of $35,000 or higher. For self-reported general health, 23.7% reported "Excellent" or "Very good," 33.6% reported "Good," and 42.8% reported "Fair" or "Poor." For the PGIS, 47% rated their lung cancer symptoms as "Not severe," 27% as "Mildly severe," and 20% as "Moderately severe."

At the time of study enrollment, 126 participants (83%) were NSCLC Stage IV and 26 (17%) were Stage IIIB. At initial diagnosis, 67% were Stage IV. About 33% of the participants were treatment naïve and half (51%) had an ECOG performance status of 1. One-hundred three participants (68%) had histologic evidence of adenocarcinoma, and 35 (23%) had histologic evidence of squamous cell carcinoma. Sixty-five participants (43%) had a comorbid clinical diagnosis of COPD and 35 (23%) were current smokers, whereas 93 (61%) were exsmokers. Table 3 shows the treatment status of the participants in the study starting with the current treatment (at the time of enrollment). The most prevalent (59%) was systemic treatment alone, followed by 7% having systemic plus radiation treatment, and 2% undergoing radiation alone.

NSCLC-SAQ descriptive characteristics

Mean item scores ranged from 0.8 for pain in chest to 2.1 for both low energy and tire easily using a response scale from 0 ("Not at all" or "Never") to 4 ("Very Severe" or "Always") (see Table 2). All items had the full range (0, 1, 2, 3, and 4) of responses endorsed. All items were answered; there were no missing data. Responses of "No <symptom> at all" or "Never," indicating poten-

Table 1

Detailed description of patient-reported outcome (PRO) data analyses and interpretation

Analysis	Description
Item-to-item correlations	A reliability analysis was conducted for all item pairs, focusing on Pearson correlation coefficients >0.70 indicating potential redundancy between the items[27] within the NSCLC-SAQ.
Item-to-total correlations	A bivariate Spearman correlation was calculated for each item score against the total score (excluding the item of interest), and any item with a value <0.40[28] was examined because this indicates that it may not be sufficiently associated with the remaining items in the hypothesized scale. These analyses were conducted between the NSCLC-SAQ items and total score.
Missing data	Frequency and percentage of missing items per participant, frequency and percentage of missing data per item, number of participants with at least 1 missing item, and number and percentage of participants with no missing data were examined.
RMT	RMT analyses were used to examine the ordering of item response options and the subscale unidimensionality. The NSCLC-SAQ items were assessed for the model fit. When the Rasch model is applied to ordered response data, where successively higher scores indicate increasing levels of agreement with a particular item, as is the case with the NSCLC-SAQ items, person ability represents how strongly respondents support the NSCLC-SAQ item and item difficulty represents how easy the item is to endorse. In addition to item threshold maps, item category trace lines were used to display the probability of a person endorsing a particular response category based on their level of support for the item and the intensity or difficulty of the item. To examine the consistency of the response pattern, a person-item distribution map was used.
Factor analyses	The results of the factor analysis were used to guide the development of the scoring algorithm for the NSCLC-SAQ. Exploratory factor analyses were performed on the 7-symptom NSCLC-SAQ with standardized factor loadings of at least 0.40 considered acceptable. A confirmatory factor analysis (principal components analysis) was conducted on the final 5-symptom NSCLC-SAQ. Fit indices (comparative fit index [values >0.9 indicates satisfactory fit],[29] goodness-of-fit index [values >0.9],[29] root mean square residual [values <0.08],[29] and root mean square error of approximation [values <0.08[29]] were assessed to confirm the relationship between the observed variables and their underlying latent constructs.[30,31]
Internal consistency reliability	Internal consistency reliability addresses the extent to which individual items within each scale are related to each other[32] and is assessed by calculating the Cronbach's alpha statistic. The values are presented descriptively on an interval level scale ranging from 0 to 1.0, with higher scores indicating a more reliable (homogeneous) instrument. Values >0.70 are generally considered indicative of a sufficiently internally consistent scale.[27] This analysis was conducted using both the 7 items and the 5 domains of the NSCLC-SAQ.
Reproducibility	The evaluation of test-retest reliability of the NSCLC-SAQ total score was made using the intraclass correlation coefficient using a 2-way mixed effect model with absolute agreement for single measures.[33,34] These analyses were conducted using the Day 1 and Day 8 data and were restricted to the subset of participants who reported that their symptoms remained stable during the study period, as defined by no change in the PGIS between day 1 and day 8.
Convergent validity	Convergent validity (demonstrating that different measures of the same concept substantially correlate when assessed concurrently) was evaluated by examining magnitude of correlations between the NSCLC-SAQ (items and total score) and the FLSI-17 (items, total score and Disease-related Symptom-Physical score). It was hypothesized that Spearman correlation coefficients of substantial magnitude (>0.40) would be apparent between the NSCLC-SAQ item and total scores and the FLSI-17 Disease-related Symptom-Physical and total scores.
Known-groups validity	Known-groups validity is the extent to which scores from a measure can discriminate between groups of participants that differ on a known relevant dimension, such as a measure/assessment of disease severity.[35] The known-groups validity of the NSCLC-SAQ was examined by grouping subjects into varying levels of disease severity/status based on the PGIS, patient self-report of general health, and clinician-reported performance status. The ability of the NSCLC-SAQ total score to discriminate between the groups of subjects according to group status was assessed via ANOVA where the PRO measure of interest was entered in the model as the dependent variable, and the known-groups variable was entered as the independent variable.

FLSI-17 = Functional Assessment of Cancer Therapy (FACT) Lung Symptom Index-17; NSCLC-SAQ = Non–Small Cell Lung Cancer Symptom Assessment Questionnaire; PGIS = Patient Global Impression of Severity; RMT = Rasch Measurement Theory.

tial ceiling effects (where participants cannot get any better), were seen most commonly in both pain items: pain in chest (51%) and pain in areas other than chest (37%). No floor effects were observed.

The 2 fatigue-related items (low energy and tire easily) had a large item-to-item correlation ($r = 0.84$) indicating redundancy (Table 4). The 2 pain items had a correlation of 0.46. Item-to-total correlations also show a strong association between each item against the rest of the items as a total score (excluding that item) other than for pain in areas other than chest (0.38). All other correlations were above 0.40.

The 2 pairs of items representing pain and fatigue were examined more closely because a unidimensional scale was preferred with none of the 5 symptom domains (ie, cough, pain, dyspnea, fatigue, and appetite) weighted more heavily in the total symptom score than the other symptom domains. Hence, to account for this and to minimize the local dependence caused by including multiple items in the overall score, the decision was made to combine the 2 items for each concept in the provisional scoring algorithm.

Fatigue

The 2 items were seen by many participants as distinct but related concepts during the qualitative research[15]; however, given the high correlation between the 2 items (0.84), indicating considerable conceptual redundancy, a score was derived by taking the mean of the 2 items, thus becoming a single fatigue score.

Pain

The two items were observed to be conceptually distinct in both the qualitative research[15] and the current study (correlation = 0.46). These 2 items individually exhibit high ceiling effects (participants indicating "No pain at all"); however, only 43 (28%) participants indicated "No pain at all" for both pain items. Therefore, because it is most clinically relevant to assess worst pain, wherever it manifests, a score was derived by taking the most severe response to either of the items, yielding a single pain score.

Factor analysis

Upon evaluating initial exploratory factor models using all 7 items and taking into consideration the overweighting of pain and fatigue (2 items each), testlet scores were created as stated above. Using a principal components analysis including the 5 domains (ie, cough, shortness of breath, poor appetite, derived pain, and derived fatigue), a single component was derived (factor loadings ranging between 0.55 and 0.77). Fit indices were adequate: comparative fit index (0.96), goodness-of-fit index (0.97), root mean

Table 2
Demographic and clinical characteristics (N = 152).

Variable	Result
Age, y	
Mean (SD)	64.3 (9.8)
Median (range)	64 (41–85)
Age*, y	
41–56	38 (25.0)
57–63	35 (23.0)
64–71	41 (27.0)
72–85	38 (25.0)
Sex*	
Female	86 (56.6)
Male	66 (43.4)
Ethnicity*	
Hispanic or Latino	8 (5.3)
Not Hispanic or Latino	144 (94.7)
Race*	
White	132 (86.8)
Black or African American	12 (7.9)
Asian	3 (2.0)
Other	5 (3.3)
Marital status*	
Married or living as married	92 (60.5)
Widowed	21 (13.8)
Separated	4 (2.6)
Divorced	24 (15.8)
Never married	11 (7.2)
Highest level of education completed*	
Less than high school	24 (15.8)
High school graduate	55 (36.2)
Some college	39 (25.6)
College graduate	25 (16.4)
Graduate or professional school	9 (5.9)
Employment status*	
Employed full-time for wages	20 (13.2)
Employed part-time for wages	6 (3.9)
Self-employed	8 (5.3)
Out of work <1 y	3 (2.0)
Out of work >1 y	7 (4.6)
Homemaker	6 (3.9)
Student	1 (0.7)
Retired	74 (48.7)
Unable to work	27 (17.8)
Annual household income from all sources*	
<$15,000	23 (15.2)
$15,000–$34,999	35 (23.0)
$35,000–$49,999	19 (12.5)
≥$50,000	63 (41.4)
Missing	12 (7.9)
Self-reported health status*	
Excellent	7 (4.6)
Very good	29 (19.1)
Good	51 (33.6)
Fair	41 (27.0)
Poor	24 (15.8)
Patient Global Impression of Severity*	
Not severe	72 (47.4)
Mildly severe	41 (26.9)
Moderately severe	31 (20.4)
Very severe	5 (3.3)
Extremely severe	3 (2.0)
NSCLC-SAQ	
1. How would you rate your coughing at its worst...?†	1.05 (0.89) [0-4]
No coughing at all*	42 (27.6)
Mild*	72 (47.4)
Moderate*	28 (18.4)
Severe*	8 (5.3)
Very severe*	2 (1.3)
2. How would you rate the worst pain in your chest...?†	0.84 (1.06) [0-4]
No pain at all*	77 (50.7)
Mild*	41 (27.0)
Moderate*	20 (13.2)
Severe*	10 (6.6)
Very severe*	4 (2.6)
3. How would you rate the worst pain in areas other than your chest...?†	1.22 (1.20) [0-4]
No pain at all*	56 (36.8)
Mild*	38 (25.0)
Moderate*	33 (21.7)

Table 2 (continued)

Variable	Result
Severe*	18 (11.8)
Very severe*	7 (4.6)
4. How often did you feel short of breath during usual activities...?†	1.81 (1.20) [0-4]
Never*	26 (17.1)
Rarely*	34 (22.4)
Sometimes*	49 (32.2)
Often*	29 (19.1)
Always*	14 (9.2)
5. How often did you have low energy...?†	2.14 (1.11) [0-4]
Never*	8 (5.3)
Rarely*	40 (26.3)
Sometimes*	46 (30.3)
Often*	38 (25.0)
Always*	20 (13.2)
6. How often did you tire easily...?†	2.14 (1.07) [0-4]
Never*	12 (7.9)
Rarely*	28 (18.4)
Sometimes*	53 (34.9)
Often*	45 (29.6)
Always*	14 (9.2)
7. How often did you have a poor appetite...?†	1.47 (1.27) [0-4]
Never*	47 (30.9)
Rarely*	32 (21.1)
Sometimes*	36 (23.7)
Often*	28 (18.4)
Always*	9 (5.9)
NCCN/FACT Lung Symptom Index-17 item†	
Total, possible range 0–68‡	22.3 (11.5) [1–50]
Disease Related Symptoms-Physical, possible range 0–40‡	14.2 (7.8) [0–33]
Disease Related Symptoms-Emotional, possible range 0–4‡	1.9 (1.3) [0–4]
Treatment Side Effects, possible range 0–12‡	2.4 (2.3) [0–9]
Function/Well-Being, possible range 0–12‡	3.8 (2.7) [0–11]
Current NSCLC stage*	
IIIB	26 (17.1)
IV	126 (82.9)
Stage at initial NSCLC diagnosis*	
I	9 (5.9)
II	3 (2.0)
III	38 (25.0)
IV	102 (67.1)
Years since initial NSCLC diagnosis	
Mean (SD)	1.1 (1.5)
Median [range]	0.5 [0.0–9.6]
Treatment status*	
Naïve	50 (32.9)
First line	49 (32.2)
Second line	26 (17.1)
Third line	27 (17.8)
ECOG performance status*	
0	49 (32.2)
1	78 (51.3)
2	25 (16.5)
Clinical diagnosis of COPD*	
No	87 (57.2)
Yes	65 (42.8)
Histological evidence of*	
Adenocarcinoma	111 (73.0)
Squamous cell carcinoma	36 (23.7)
Unknown	5 (3.3)
Mutation test and status*	
EGFR mutation+	14 (9.2)
ALK+ negative	23 (15.1)
EGFR mutation+ and ALK+ negative	3 (2.0)
None of the above	53 (34.9)
Not tested	57 (37.5)
Missing	2 (1.3)
Smoking history*	
Current smoker	35 (23.0)
Exsmoker	93 (61.2)
Never a regular smoker	23 (15.1)

ALK = anaplastic lymphoma kinase; COPD = chronic obstructive pulmonary disease; ECOG = Eastern Cooperative Oncology Group; EGFR = epidermal growth factor receptor; NCCN/FACT = National Comprehensive Cancer Network/Functional Assessment of Cancer Therapy Lung Symptom; NSCLC = non–small cell lung cancer; NSCLC-SAQ = Non–Small Cell Lung Cancer Symptom Assessment Questionnaire.

* Values are presented as n (%).

† Values are presented as mean (SD) [range].

‡ Higher scores indicate a severely symptomatic patient.

Table 3
Treatment (Tx) status (current and history) (N = 152).

Tx*	First-line Tx	Second-line Tx	Third-line Tx	Subsequent Tx
Current Tx at time of enrollment				
Radiation alone	2 (1.3)	1 (0.7)	–	–
Systemic Tx alone	42 (27.6)	26 (17.1)	21 (13.8)	
Radiation and systemic	9 (5.9)	1 (0.7)	–	–
No current Tx†	–	–	–	–
Tx received				
Surgery	3 (2.0)	4 (2.6)	–	–
Radiation	6 (3.9)	5 (3.3)	1 (0.7)	–
Systemic Tx	87 (57.2)	49 (32.2)	23 (15.1)	10 (6.6)
Surgery + systemic Tx	3 (2.0)	–	–	–
Radiation + systemic Tx	21 (13.8)	2 (1.3)	1 (0.7)	1 (0.7)
Surgery + radiation + systemic Tx	3 (2.0)	–	–	–
Not applicable	29 (19.1)	92 (60.5)	127 (83.6)	–
Those who received systemic Tx				
Received a platinum-based regimen	101 (66.4)	16 (10.5)	4 (2.6)	2 (1.3)
Received a targeted therapy	35 (23.0)	31 (20.4)	13 (8.6)	5 (3.3)

* Values are presented as n (%).
† Fifty patients (32.9%) were not currently undergoing Tx.

square residual (0.05) and root mean square error of approximation (0.08). When evaluating by treatment group (treatment naïve versus currently treated with systemic and/or radiation), no differences were observed.

RMT analysis

RMT analyses allows for examination of the ordering of item response options and the scale unidimensionality. The item threshold map shows that all 7 items were correctly ordered; that is, the threshold values between adjacent pairs of response options are ordered by magnitude. The items' response categories reflect an ordered continuum from "No <symptom> at all" to "Very severe <symptom>" (items 1–3) or "Never" to "Always" (items 4–7), where each response had its own probability of being adequately endorsed. Responses of 0 "No/Never" are independent of the response 1 "Mild/Rarely," in turn independent of response 3 "Moderate/Sometimes," and so on. The distributions of person and item threshold locations for the NSCLC-SAQ showed that the items covered the range of persons included and that both the items and participants were reasonably well distributed.

NSCLC-SAQ scoring

The provisional scoring algorithm of the NSCLC-SAQ total score is as follows:

Cough domain score

Score of the cough item, or missing if skipped.

Fatigue domain score

If both items present, compute mean; or use score from 1 item if the other is missing; or missing if both are skipped.

Pain domain score

If both items present, use most severe of both; or use score from 1 item if the other is missing; or missing if both are skipped.

Dyspnea domain score

Score of the shortness of breath item, or missing if skipped.

Appetite domain score

Score of the poor appetite item, or missing if skipped.

NSCLC-SAQ total score

Sum all 5 domain scores; if any are missing, a total score is not computed. This creates a total score ranging between 0 and 20, with higher scores indicating more severe symptomatology.

Reliability

Internal consistency reliability was examined using Cronbach α for the NSCLC-SAQ 7 items (0.78) and 5 domains (0.72). The evaluation of test–retest reliability was conducted using the intraclass correlation coefficient (ICC). These analyses were restricted to the subset of patients whose NSCLC-related symptom status remained stable during the study period as defined by providing the same response to the PGIS on day 1 and day 8. Of the 148 patients who completed a retest within the acceptable window, 90 (60.8%) provided the same PGIS responses on day 1 and day 8. The ICC was 0.87 with a 95% CI of 0.80 to 0.91. As a post hoc analysis, the PGIS change definition was expanded to allow a 1-point change from day 1 to day 8. Of the 148 patients who completed the retest, 133 (89.8%) had no change or only a 1-point change in PGIS. ICC values were 0.82 (95% CI, 0.76–0.87).

Construct validity assessment

Convergent validity was assessed by examining the magnitude of correlations between the NCSLC-SAQ items and the FLSI-17 items. All associations hypothesized to have stronger correlations (>0.40) between items of the NSCLC-SAQ and FLSI-17 were met (see Table 4).

Known-groups validity was examined using the PGIS, self-reported health status, and ECOG performance status. The NSCLC-SAQ total score was able to differentiate between levels of self-reported symptom severity on the PGIS (not severe, mildly severe, moderately severe, or very/extremely severe; $P < 0.001$), self-reported health status (excellent, very good, good, fair, or poor, $P < 0.001$), and clinician-reported performance status (ECOG 0, ECOG 1, and ECOG 2; $P < 0.001$) (see Figure 2).

Table 4

Non–Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ) correlations (item-to-item, item-to-total, and by Functional Assessment of Cancer Therapy Lung Symptom Index-17 [FLSI-17] items) (n = 152).

Variable	1. Cough	2. Chest pain	3. Other pain	Pain score (worst)	4. Shortness of breath	5. Low energy	6. Tire easily	Fatigue score (mean)	7. Poor appetite	NSCLC-SAQ Total
NSCLC-SAQ item										
1. Cough	—									.412**
2. Chest pain	.297**	—								.413**
3. Other pain	.171*	.455**	—							.381**
Pain score (worst)	.226**	.641**	.907**	—						.357**
4. Shortness of breath	.410**	.152	.136	.178*	—					.476**
5. Low energy	.294**	.173*	.326**	.324**	.460**	—				.664**
6. Tire easily	.251**	.216**	.283**	.326**	.457**	.844**	—			.664**
Fatigue score (mean)	.288**	.194*	.307**	.327**	.473**	.964**	.954**	—		.580**
7. Poor appetite	.383**	.326**	.303**	.354**	.382**	.481**	.458**	.489**	—	.576**
FLSI-17										
1. I have a lack of energy	.333**	.278**	.358**	.392**	.447**	.764**	.754**	.790**	.521**	.721**
2. I have pain	.168*	.643**	.710**	.774**	.213**	.411**	.453**	.449**	.401**	.598**
3. I am losing weight	.181*	.244**	.204*	.268**	.183*	.336**	.292**	.327**	.596**	.465**
4. I have been short of breath	.325**	.209**	.151	.204*	.853**	.495**	.515**	.525**	.371**	.666**
5. I feel fatigued	.269**	.261**	.356**	.390**	.433**	.764**	.746**	.786**	.537**	.706**
6. I have been coughing	.872**	.270**	.200*	.215**	.376**	.256**	.253**	.264**	.365**	.575**
7. I have bone pain	.088	.436**	.579**	.568**	.108	.260**	.243**	.262**	.266**	.389**
8. Breathing is easy for me	.243**	.263**	.303**	.343**	.582**	.390**	.448**	.435**	.354**	.576**
I feel tightness in my chest	.290**	.420**	.300**	.365**	.421**	.305**	.340**	.335**	.246**	.482**
9. I have a good appetite	.279**	.323**	.350**	.397**	.326**	.423**	.390**	.423**	.827**	.674**
10. I am sleeping well	.109	.254**	.272**	.307**	.218**	.324**	.406**	.379**	.330**	.399**
11. I worry that my condition will get worse	.009	.189*	.141	.178*	.111	.071	.003	.039	.255**	.186*
12. I have nausea	.119	.304**	.262**	.331**	.241**	.451**	.426**	.457**	.424**	.468**
13. I am bothered by hair loss	.105	.101	.096	.088	.121	-.006	-.043	-.025	.100	.115
14. I am bothered by side effects of [Tx]	.140	.210**	.254**	.347**	.220**	.238**	.247**	.252**	.311**	.378**
15. My thinking is clear	-.063	.134	.157	.205	.002	.166	.113	.146	.127	.131
16. I am able to enjoy life	.114	.178*	.265**	.290**	.384**	.491**	.390**	.459**	.454**	.508**
17. I am content with the quality of my life right now	.164*	.267**	.327**	.376**	.353**	.422**	.330**	.392**	.532**	.544**
FLSI-17 DRS-P	.426**	.483**	.522**	.581**	.573**	.650**	.622**	.662**	.708**	.872**
FLSI-17 Total score	.345**	.470**	.515**	.587**	.531**	.645**	.617**	.656**	.701**	.833**

DRS-P = Disease-Related Symptoms–Physical; Tx = treatment.

* Correlation is significant at the 0.05 level (2-tailed).

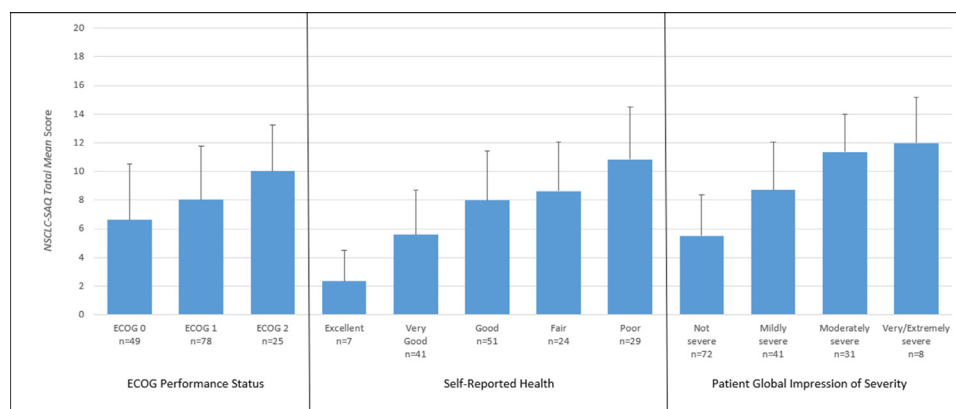** Correlation is significant at the 0.01 level (2-tailed).

Figure 2. Evidence for known groups validity of the Non–Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ). Lower scores indicate lower symptom severity. Overall significance for all comparisons was $P < 0.001$. ECOG = Eastern Cooperative Oncology Group.

Discussion

In regard to oncology, the FDA has made it clear they are interested in reviewing clinical trial data that include assessment of the following core PROs: symptomatic adverse events, physical function, and disease-related symptoms.[3] The NSCLC-SAQ was designed to assess NSCLC-related symptoms in clinical trials in a well-defined and reliable way. Although several measures aimed at assessing patient-reported NSCLC-related symptoms had been previously developed (eg, FACT-L, M.D. Anderson Symptom Inventory Lung Cancer Module, Lung Cancer Symptom Scale, and European Organization for Research and Treatment of Cancer Quality of Life Lung-Specific Questionnaire), it was not clear to the NSCLC Working Group that sufficient evidence documenting the provenance of these legacy measures could be assembled to meet the evidentiary expectations of the FDA's qualification program.[8] With the release of FDA's PRO Guidance and the increased focus on the use of rigorously developed PRO measures as clinical trial end point measures, ensuring the adequacy of symptom inventories used to support labeling claims necessitates a structured review of evidence supporting content validity and the psychometric properties of these existing instruments. As such, a new NSCLC symptom measure was developed for the specific purpose of capturing a symptom-based efficacy endpoint in clinical trials for advanced NSCLC. The authors do note that more recent statements from FDA indicate a greater openness to the qualification of legacy PRO measures than when the NSCLC Working Group's qualification project began in 2012.[24]

As recommended by FDA, the use of a mixed-methods approach (using both qualitative and quantitative information) and the early use of quantitative data to further support the content validity of items and scales is a prudent and productive approach to PRO measure development.[10–14] The primary aim of this cross-sectional observational study was to evaluate the performance of the NSCLC-SAQ, both on an individual item level and scale level. The NSCLC-SAQ was psychometrically tested using both classical as well as modern analyses (ie, RMT).

Rasch analyses showed that the items were ordered, and the person-to-item distribution was good. Factor analysis indicated a single component supporting the use of a single (total) score. The 2 pain items (worst response) and the 2 fatigue items (mean of responses) are combined to create single item scores. Internal consistency of the NSCLC-SAQ was acceptable ($\alpha = 0.78$) and test–retest reproducibility was good with an ICC of 0.87. Convergent validity was supported as the NSCLC-SAQ score was substantially correlated (0.87) with the FLSI-17 DRS-P score. The NSCLC-SAQ differentiated between levels of self-reported symptom severity (ie, PGIS), clinician-reported performance status (ie, ECOG), and self-reported

health status ($P < 0.001$). We acknowledge this study's limitation with respect to the severity of this sample; small numbers of participants were in the very and extremely severe groups. This will need to be investigated further to make more accurate comparisons within these more severe groups. In terms of the scoring, additional exploration/confirmation with data from interventional clinical studies is warranted around the use of the fatigue and pain testlets. Further empirical evidence may lead to the elimination of 1 of the fatigue items.

An additional limitation was that this study included only US patients and only those who spoke English; however, a formal translatability assessment[25] was conducted to optimize the NSCLC-SAQ item language to facilitate future translation and cultural adaptation through early identification of potential difficulties.[15]

In addition, the NSCLC-SAQ's longitudinal measurement properties need to be evaluated. A key next step for the NSCLC-SAQ is to examine its ability to detect meaningful change within advanced NSCLC treatment trials. Now that the NSCLC-SAQ has obtained FDA qualification[26] it is publicly available. Sponsors of advanced NSCLC clinical trials are encouraged to incorporate it into their PRO measurement strategy in early-phase studies to help build evidence for its performance before being used as part of a primary or secondary efficacy end point in confirmatory trials.

Conclusions

The cumulative evidence on content validity, construct validity, and reliability of the NSCLC-SAQ, including the quantitative study described above, led to its qualification by FDA as a drug development tool in a limited context of use. The qualification supports the NSCLC-SAQ as a patient-reported measure of symptoms in advanced NSCLC drug development.[26] Further evaluation is needed regarding the NSCLC-SAQ's longitudinal measurement properties (eg, sensitivity to change and responsiveness) and the interpretation of clinically meaningful within-patient score change. Implementing assessment with the NSCLC-SAQ across sponsors will, ultimately, enable comparison of advanced NSCLC treatment trial results and facilitate comparative effectiveness research by providing a standard measure of patient-reported clinical benefit.

Acknowledgments

Conflicts of Interest

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*. 2018;68:7–30.
2. Rivera MP, Mehta AC, Wahidi MM. Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*. 2013;143:e142S–e165S.
3. Kluetz PG, Slagle A, Papadopoulos EJ, Johnson LL, Donoghue M, Kwitkowski VE, Chen W-H, Sridhara R, Farrell AT, Keegan P, Kim G, Pazdur R. Focusing on core patient-reported outcomes in cancer clinical trials: symptomatic adverse events, physical function, and disease-related symptoms. *Clinical Cancer Research (Clin Cancer Res)*. 2016;22(7):1553–1558.
4. Cella DF, Bonomi AE, Lloyd SR, et al. Reliability and validity of the Functional Assessment of Cancer Therapy-Lung (FACT-L) quality of life instrument. *Lung Cancer*. 1995;12:199–220.
5. Bergman B, Aaronson N, Ahmedzai S, et al. The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials. *European Journal of Cancer*. 1994;30:635–642.
6. Hollen PJ, Gralla RJ, Kris MG, Kris MG. Measurement of quality of life in patients with lung cancer in multicenter trials of new therapies. Psychometric assessment of the Lung Cancer Symptom Scale. *Cancer*. 1994;73:2087–2098.
7. Mendoza TR, Wang XS, Lu C, et al. Measuring the symptom burden of lung cancer: the validity and utility of the lung cancer module of the M. D. Anderson Symptom Inventory. *The Oncologist*. 2011;16:217–227.
8. Food Drug Administration (FDA). Guidance for Industry and FDA Staff: Qualification Process for Drug Development Tools. January 2014. Available from: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm230597.pdf. [Accessed March 3, 2019].
9. Coons SJ, Kothari S, Monz BU, et al. The patient-reported outcome (PRO) consortium: filling measurement gaps for PRO end points to support labeling claims. *Clinical Pharmacology and Therapeutics*. 2011;90:743–748.
10. Food Drug Administration (FDA)Guidance for Industry Patient-reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. *Fed Regist*. 2009;74:65132–65133.
11. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1–eliciting concepts for a new PRO instrument. *Value Health*. 2011;14:967–977.
12. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2–assessing respondent understanding. *Value Health*. 2011;14:978–988.
13. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health*. 2007;10(Suppl 2):S125–S137.
14. Rothman M, Burke L, Erickson P, et al. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. *Value Health*. 2009;12:1075–1083.
15. McCarrier KP, Atkinson TM, DeBusk KP, et al. Qualitative Development and Content Validity of the Non-small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ), A Patient-reported Outcome Instrument. *Clinical therapeutics*. 2016;38:794–810.
16. US National Cancer Institute (NCI). Common Terminology Criteria for Adverse Events (CTCAE) Version 4.03. June 14, 2010. Available from: http://evs.nci.nih.gov/ftp1/CTCAE/About.html. [Accessed March 3, 2019].
17. Yount S, Beaumont J, Rosenbloom S, et al. A brief symptom index for advanced lung cancer. *Clin Lung Cancer*. 2012;13:14–23.
18. Nunnally JC, Bernstein IH. *Psychometric theory*. McGraw-Hill; 1994.
19. Hambleton RK, Slater SC. Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*. 1997;13:21–28.
20. SPSS Inc. *PASW Statistics for Windows*. 180 ed. Chicago: SPSS Inc; 2009.
21. Andrich D, Sheridan BS, Luo G. *RUMM2030: Rasch unidimensional measurement models [software]*. Perth, Australia: RUMM Laboratory; 2012.
22. Arbuckle J. *IBM SPSS AMOS 20 for Windows*. Chicago, IL: SPSS; 2011.
23. Byrne BM. *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. 2nd ed. Multivariate applications series. New York, NY: Routledge/Taylor & Francis Group; 2010.
24. Papadopoulos E, Bush EN, Eremenco S, Coons SJ. Why reinvent the wheel? Use or modification of existing clinical outcome assessment tools in medical product development. *Value in Health*. 2019 Article in press (published online October 16. doi:10.1016/j.jval.2019.09.2745.
25. Acquadro C, Patrick DL, Eremenco S, et al. Emerging good practices for Translatability Assessment (TA) of Patient-Reported Outcome (PRO) measures. *J Patient Rep Outcomes*. 2017;2:8.
26. Food Drug Administration (FDA). Qualification of Non-Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ) – A Patient-Reported Outcome Instrument. April 4, 2018. Available from: https://www.fda.gov/media/119250/download. [Accessed January 22, 2021].
27. Carmines EG, Zeller RA. *Reliability and Validity Assessment*. SAGE Publications; 1979.
28. Gliem J, Gliem R. *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Columbus, OH: Ohio State University; 2003.
29. Hair J, Black W, Babin B, et al. *Multivariate data analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall; 2010.
30. Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 1999;6:1–55.
31. Kline RB. Principles and Practice of Structural Equation Modeling, *Fourth Edition*. Guilford Publications; 2015.
32. Cronbach LJ, Warrington WG. Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika*. 1951;16:167–188.
33. McGraw KO, Wong SP. Forming inferences about some intraclass correlations coefficients. *Psychological Methods*. 1996;1:30–46.
34. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979;86:420–428.
35. Hays RD, Revicki D. *Assessing Quality of Life in Clinical Trials: Methods and Practice*. 2nd ed. Reliability and validity (including responsiveness). New York, NY: Oxford University Press; 2005.