VALUE IN HEALTH 14 (2011) 291–306

available at www.sciencedirect.com

SCIENCE DIRECT®

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

Development of a Conceptual Framework and Calibrated Item Banks to Measure Patient-Reported Dyspnea Severity and Related Functional Limitations

Seung W. Choi, PhD[a,], David E. Victorson, PhD[a], Susan Yount, PhD[a], Susan Anton, DrPh[b], David Cella, PhD[a]*

[a] *Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[b] *Boehringer Ingelheim Pharmaceuticals, Inc., Medical Affairs, Ridgefield, CT, USA*

Keywords:
Dyspnea
COPD
Scale development
Measurement
Computer adaptive testing

A B S T R A C T

Objectives: Chronic obstructive pulmonary disease is a major global health problem. Although several patient-reported outcome (PRO) measures of chronic obstructive pulmonary disease exist, none were developed using patient-driven concept development. We developed an item bank for dyspnea severity and related functional limitations on the basis of a PRO conceptual framework derived from patient input.

Methods: We identified a large pool of existing items based on a conceptual framework and literature review. Using patient and expert review panels and an item refinement/modification process, we developed an item bank aligned with the conceptual framework, which subsequently underwent psychometric testing via an online Internet panel of dyspnea patients (N = 608).

Results: Exploratory factor analysis suggested a dominant first factor accounting for about 78% of the total variance. Confirmatory factor analysis supported a unidimensional model. Item response theory analysis demonstrated good model fit, and differential item functioning analyses indicated that the 33-item scale showed potential for measurement equivalence across sex. A 10-item short form produced comparable scores ($r = 0.98$) and a computerized adaptive-testing simulation indicated efficient measurement with fewer items (mean 4.65 items).

Conclusions: An efficient patient-reported measure of dyspnea severity and related functional limitations, based on a patient-driven PRO conceptual framework, is now available for further validation and use.

Funding: This work was supported in part by a grant from Boehringer Ingelheim Pharmaceuticals.

* *Address correspondence to:* Seung W. Choi, PhD, Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 710 North Lake Shore Drive, Chicago, IL 60611, USA.
E-mail address: s-choi@northwestern.edu.

1098-3015/$36.00 – see front matter
doi:10.1016/j.jval.2010.06.001

Introduction

Chronic obstructive pulmonary disease (COPD) is a major global health problem[1] and can significantly disrupt the daily lives of those who live with it [2]. One of the primary symptoms of COPD is dyspnea, or shortness of breath (SOB), which results from disease-related airflow obstruction [3]. Similar to the measurement of pain or fatigue, SOB can be challenging to evaluate because it can be evaluated only from the individual's subjective perspective and its reporting can be complicated by cognitive, emotional, and other factors.

The US Food and Drug Administration (FDA) has drafted a guidance document for measuring patient-reported outcomes (PROs) when seeking a label claim [4]. A well-constructed and documented PRO instrument can facilitate interaction with regulatory agencies such as the FDA and the European Medicines Agency. The standards in the FDA guidance place special attention on methodologic and procedural steps that are important during scale development such as direct patient input (usually through focus groups or interviews) into concept identification for subsequent items. In addition, the guidance underscores the role of a PRO conceptual framework that articulates expected relationships between items within domains and the larger instrument, as well as a trial-specific endpoint model (relationships of concepts used as endpoints to support a label claim) [4]. Although several PRO measures of SOB exist, each has its limitations and none has been developed on the basis of patient-driven concept development.

In addition to growing adherence to PRO development guidance from regulatory agencies, another transformation has been occurring in the field of patient reported health outcomes assessment—the use of modern test development approaches, such as item response theory (IRT) [5,6]. A member of a larger family of mathematical models, IRT is used to "calibrate" patient item responses along a severity continuum that represents some latent trait, such as SOB. Using a calibrated "item bank" allows one to estimate a person's location along a continuum of increasingly difficult items and to determine which items provide maximum information of a given concept [7,8].

We have previously reported our work leading to the development of a dyspnea-specific conceptual model [9]. At the core of that model are the related and interacting concepts of dyspnea and functional limitation (FL). We now describe the steps we took to develop and test an FDA-guidance based PRO conceptual framework and an IRT-based item bank of dyspnea severity and related FLs.

Methods

Figure 1 provides an overview of these steps; the following sections provide greater detail on Phases II, IV, and V. A comprehensive multilingual translation was also conducted according to our methodology [10]; however, report of that work is beyond the scope of this article.

Creation of a new measure

Based on concepts that were identified during the development of our dyspnea-specific conceptual model [9], we sought to locate available items and write new ones when others were not available. We used an item identification and generation process drawn from our National Institutes of Health Roadmap Patient-Reported Outcomes Measurement Information System initiative [11,12], and as described below. To inform this process, we conducted a literature search to identify existing items in published self-report measures of dyspnea and health-related quality of life in COPD. Using keywords "dyspnea," "pulmonary disease," and "chronic obstructive," we conducted a MEDLINE search from 1969 through 2004, to identify measures of dyspnea and related problems associated with pulmonary disorders. We specified English language studies only. Questionnaires were retained for further evaluation if they measured dyspnea as a primary focus or an explicit subscale. Single-item scales were excluded. Scales that used a visual analogue scale format were excluded, because the team had previously determined that flexibility in administration options was important to maintain. We checked the content of retained questions against those asked in visual analog scale format to ensure that the instruments remaining under consideration were representative of content. We also excluded interviewer-administered scales for the same reason, and similarly checked that content of such instruments was not unique to this mode of administration.

Items that were retained for the library were categorized ("binned") into conceptually similar groups of questions. These bins were used to group highly redundant questions within larger concepts. Highly redundant items were thus clustered together to enable selection of the most desirable components relating to description of the item context (e.g., time frame, circumstances, conditional statements), linguistic expression of the item stem, and response options.

The 15 patients who were interviewed to inform the conceptual model [9] were also interviewed to rate how they evaluated everyday activities common to many existing questionnaires. This component of the interview was conducted after seeking input regarding basic concepts. Specifically, patients were asked, on a zero to three scale, how important items reflecting a wide array of activities were to them (zero = not at all relevant; three = extremely relevant). A cutoff (mean rating < 1.0) was used to flag potentially unimportant items. During this item review process, additional codes were applied to items to inform item modification and removal decisions. "Low item rating" was coded when patients rated an item as unimportant (mean $<$ 1.0). "Prevalence" was coded when activities described within items were rated as uncommon. "Misfit" was coded when an item performed poorly in preliminary psychometric analyses of archival data. "Sex bias" was coded when an activity described in an item appeared to apply exclusively or primarily to one sex over another. "Translatability" was coded when an item's English version had characteristics or words known to cause problems when translating into another language. "Cultural bias" was coded when an activity

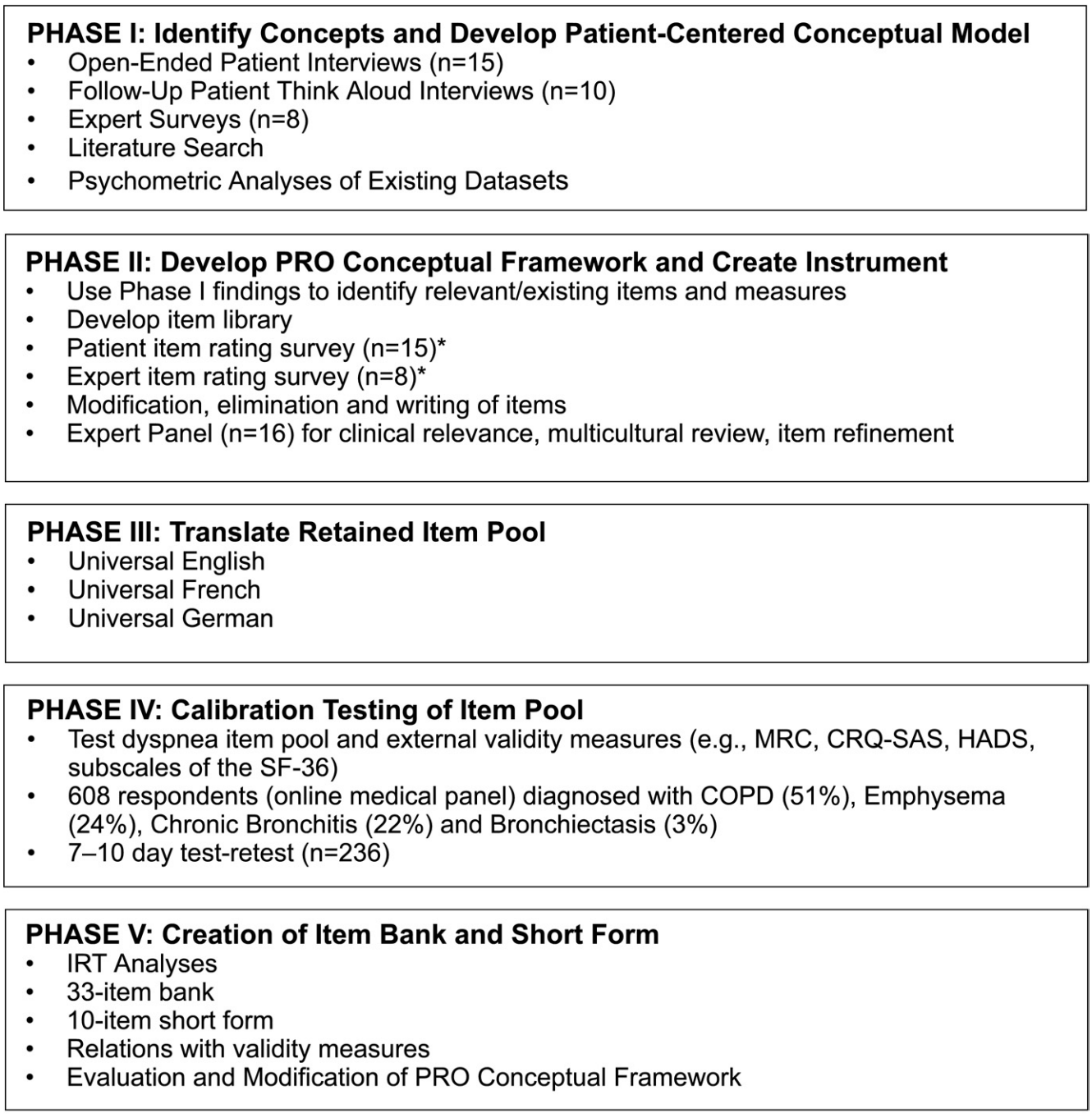


Fig. 1 – Steps taken to develop a new measure of dyspnea severity and related functional limitations. PRO, patient-related outcomes. MRC, Medical Research Council Dyspnea Scale. CRQ-SAS, Chronic Respiratory Questionnaire–Self-Administered. HADS, Hospital Anxiety and Depression Scale. COPD, chronic pulmonary obstructive disease.

described in an item appeared to be biased toward culture of the United States. "Impracticality of activity" was coded when an activity described in an item seemed highly impractical for a person with COPD. Finally, "Construction" was coded for poorly written items that were removed (e.g., multibarreled, incorrect grammar). In line with recommended guidelines for data saturation determination [12], we considered concepts "saturated" when at least 70% of the sample mentioned them.

In addition to obtaining patient input, two different expert review groups were assembled: a clinical review panel composed of clinical experts in pulmonary and respiratory medicine (n = 8), and an item review panel, composed of outcomes researchers, health psychologists, psychometricians, and ex-

perts in translation science (n = 16). Our clinical review panel rated how important certain activities were to their patients, using the same zero to three scale (zero = not at all relevant; three = extremely relevant). A cutoff was used to flag potentially unimportant items (mean < 1.0). Finally, expert ratings were compared to patient ratings.

Following this initial review, highly redundant items within bins were sorted by the item review panel, whose members selected from among the available items in each bin those that offered the clearest expression of the content. Study and discussion of bins at this point sometimes led to item wording modification, and several new items were written to fill conceptual gaps. After a complete round of winnowing, the item review panel re-examined items for redundancy, clarity, and translatability.

Patient sample for item calibration and scaling

To maximize geographic diversity as well as efficiency of data collection, calibration testing was conducted using an online Internet panel of COPD patients across the United States. We collected Internet-based response data from 608 individuals at baseline and a subset of 236 respondents at 7 to 10 days' follow up. Respondents were sampled from an online panel testing company and were matched against a frame of records randomly selected from the US-representative 2004 American Community Survey. To ensure a range of dyspnea severity, COPD patients on the Internet panel were enrolled based on their response to a screening question that assigned them into one of the Medical Research Council (MRC) classifications [14]: zero = no breathlessness; one = I only became breathless after strenuous exercise; two = I only became breathless when hurrying on level ground or walking up a slight hill; three = I had to walk slower than other people on level ground because of breathlessness or I had to stop for breath even when walking at my own pace; four = I had to stop for breath after walking about 100 yards (or after a few minutes) on level ground; five = I was too breathless to leave the house, or breathless when dressing or undressing. Internet response data were used for the scaling and item calibration analyses. The goal was to have no single category with more than 30% of patients, and at least 10% at category 4 or worse. Actual MRC scores were as follows: 29% = 1; 30% = 2; 27% = 3; 10% = 4; 3% = 5.

The majority of the sample was male (60%; n = 362), white (96%; n = 585), and married (58%; n = 350). Most respondents reported completing some college or attaining a college degree (41% and 39%, respectively). Self-reported diagnoses included: COPD (n = 313), emphysema (n = 146), chronic bronchitis (n = 131), and bronchiectasis (n = 18). With regard to smoking status, 26% (n = 156) acknowledged they currently smoke tobacco. A total of 75% (n = 455) reported using inhalers, and 45% (n = 203) steroid inhalers. A subset reported having their most recent exacerbation within the past month (28%; n = 170) or between 1 and 3 months (19%; n = 114). In terms of the severity of the most recent exacerbation, 60% (n = 364) reported that it was mild, 24% (n = 148) moderate, and 7% (n = 40) severe.

Dimensionality

We conducted confirmatory factor analysis (CFA) and specified a 1-factor model for items pooled and deemed to capture dyspnea symptoms. We also used exploratory factor analysis (EFA) to provide supplemental information (e.g., the magnitude of secondary dimensions) should the unidimensionality not be confirmed. We sought to determine whether the item pool is sufficiently unidimensional to be adequately modeled by a unidimensional IRT model. One challenge we faced when conducting CFA and EFA was the number of missing responses present in the data. Listwise deletion resulted in a substantial decrease in the sample size available for the analysis. For the Dyspnea bank, only 76 of 608 cases had a complete set of responses on the 33 questions. For the Functional Limitation (FL) bank, 126 of 608 cases had complete responses. Furthermore, we could not assume that responses on the dyspnea severity questions are missing at random (because respondents with more severe COPD symptoms are more likely to skip questions on more vigorous activities). Therefore, we conducted CFA only on the FL items using pair-wise deletion with all cases having valid data for each pair of items. We also conducted EFA with pair-wise deletion where each element of the matrix of polychoric correlations was estimated using all available data.

IRT modeling

IRT modeling allows us to evaluate the quality of the rating scales and the fit of the observed response data to the single underlying latent trait being measured by the collection of questions. The data from the two banks were fitted separately to the graded response model [15], using MULTILOG (version 7, Scientific Software International, Lincolnwood, IL). The graded response model allows the item discrimination (slope) parameter to vary across items.

Differential item functioning (DIF)

DIF is a major threat to the validity of test scores. DIF can affect test scores for examinees in certain demographic groups, independent of the construct being measured. One key variable of interest in the current study was sex; that is, whether patients' sex affects how they respond to the questions over and above their levels on the dyspnea severity trait being measured. We used an ordinal logistic regression technique for detecting DIF items using lordif [16]. Ordinal logistic regression provides a flexible framework for detecting various types of DIF (e.g., uniform, nonuniform). We substituted the matching variable based on sum scores with IRT-based trait scores [17]. The use of the IRT trait level score in lieu of the traditional sum score makes this approach more robust and applicable when there are many missing responses in the data.

Computerized adaptive testing (CAT) of the Dyspnea bank

Upon developing an item bank, a post-hoc CAT simulation can provide useful information regarding the potential effectiveness of the bank under CAT administrations for a population of interest. CAT enables shortening of tests by adaptively administering "optimal" items. We conducted a CAT simulation using Firestar software [18] with the standard error stopping

criterion of 0.3 and the minimum and maximum number of items to administer set at three and 12, respectively.

Results

Identification of existing items and measures

A total of 220 unique citations were obtained from the Medline literature search. About half (115) were excluded from further examination for the following reasons: 91 were case reports, comments, letters to editors, quizzes giving continuing education credits, meta-analyses, literature reviews, or qualitative reports, and 24 were empirical studies that did not identify a patient-reported measure of dyspnea. The remaining 105 articles were published in 43 different journals. From these 105 articles, 14 different questionnaires were identified and reviewed by our team of psychometric, clinical, and social science experts. Through this process, seven questionnaires were identified and retained for further review. These included: 1) Pulmonary Functional Status & Dyspnea Questionnaire - Original [19] and Pulmonary Functional Status & Dyspnea Questionnaire - Modified [20]; 2) The Breathlessness, Cough, and Sputum Scale [21]; 3) London Chest Activity Daily Living Scale [22]; 4) University of Cincinnati Dyspnea Questionnaire [23]; 5) University of California - San Diego Shortness of Breath Questionnaire [24]; 6) Chronic Respiratory Questionnaire - Self-Administered Individualized and Self-Administered Standardized (CRQ-SAS) [25]; and 7) St. George's Respiratory Questionnaire - American translation [26].

Development of an item library

These seven questionnaires contributed 364 items into the library. Of these, 104 items were excluded because they did not assess conceptual model concepts such as dyspnea and/or functional limitations (n = 40), asked more than one question (n = 27), were unclear (n = 23), or were open-ended "write-in" questions (n = 14). The remaining 260 questions were binned into conceptually similar groups of questions. These bins were intended to group highly redundant questions within larger concepts. As a result, we created many more bins than concepts, as our purpose was to cluster highly redundant items together to enable selection of the most desirable components relating to description of the item context (e.g., time frame, circumstances, conditional statements), linguistic expression of the item stem, and response options). Members of the item review panel allocated the items into 28 separate bins, which were created from a synthesis of the most commonly occurring domains and themes from available self-report dyspnea measures. These bins included: personal hygiene, dressing, raising arms/reaching, walking, uphill/climbing stairs, hurry/brief running, getting up, bending, sexual activity, angry/upset, eating/cooking/beverages/medication, doing dishes/light cleaning, heavier cleaning, laundry/making beds, picking up/carrying/moving, household chores/shopping, lawn activities, car activities, home repair, travel/recreation, social/recreation, solitary recreation, exercise/sports, talking while doing activities, at rest, emotional impact, self-efficacy, and general shortness of breath.

Following this initial classification, highly redundant items within bins were winnowed by the item review panel. Panel members selected from among the available items in each bin those that offered the clearest expression of the content. Study and discussion of bins at this point sometimes led to item wording modification, and several new items were written to fill conceptual gaps. After a complete round of winnowing, the item review panel re-examined items for redundancy, clarity, and translatability. This resulted in the 111 retained items that were included in the Patient and Expert Item Rating Interviews.

Patient and expert item rating interviews

We observed redundancy and replication of concepts within the 15 open-ended interviews, of which the majority of concepts exceeded our a priori cutoff (70%). On average, 13% of the item pool was rated "not at all relevant," 46% was rated "somewhat relevant," 32% was rated "very relevant," and 4% was rated "extremely relevant" by patients and experts. Often there was convergence, but when expert ratings diverged with patient ratings (e.g., they rated an item as less or more important) the patient ratings took precedence. There were a few occasions when an expert's high importance rating led us to include the item going forward, as we chose to be conservative about item pool reduction at this early stage. This resulted in 90 retained items.

Item refinement and review

During these item selection and refinement meetings, every attempt was made to identify and/or write new items to reflect aspects of key elements from the conceptual model. Members of our clinical and item review panels met to incorporate information to determine whether certain items or clusters of items had deeper conceptual coverage, thus being more equipped to more clearly distinguish between discrete levels of performance or limitation on a given task. This process helped to identify commonalities and discrepancies between measures and comments and identify conceptual or hierarchical gaps. This process had a significant influence on the design of the instrument and the current item pool. In the revised set of draft items, we made selected item modifications to account for this distinction and clarify a level of function in the stem. For example, we changed the item "taking a bath" to "taking a bath without assistance." This resulted in newly written items, increasing the pool to 119. Panel members then selected a bank of 50 items from the pool that pertained specifically to dyspnea. This involved a multistep, iterative process in which experts included items that covered the range of potential dyspnea-specific symptoms and issues, while winnowing out items that were redundant or not conceptually applicable. Finally, we added 50 items assessing influence of dyspnea on function, using items that mirrored the content in the 50 dyspnea items. These FL items had the same stem content, but a different context and set of response options as the 50 dyspnea items. Essentially these FL items have the same item stem content as the dyspnea items; however, a different context and set of response options are used than the 50 dyspnea items. Therefore, dyspnea items refer to severity of shortness of breath, whereas FL items deal with the

extent of limitation from dyspnea. These 100 items (50 pertaining to dyspnea severity and 50 pertaining to FL), supplemented with questions that assessed time extension, task avoidance, emotional response to dyspnea, activity requirements, assistive devices, and exposure to airborne irritants, brought the total field testing total to 169 items.

Response options, recall period, and patient understanding

Members of the clinical expert panel evaluated several different types of response options from existing scales and expert input. After discussion of the advantages and disadvantages of various options in light of the interview data and clinician input, we decided upon Likert scaling of intensity and difficulty. For dyspnea severity we selected: zero = no shortness of breath; one = mildly short of breath; two = moderately short of breath; three = severely short of breath; four = did not do this in the past 7 days. For FL we selected: zero = no difficulty; one = a little difficulty; two = some difficulty; three = much difficulty. Patient understanding of concepts was evaluated during individual think-aloud interviews (n = 10) during the conceptual-model–development phase [9].

In terms of recall period, our clinical experts unanimously and strongly recommended a time frame longer than 2 days, to capture enough patient experience with a range of functional activities and symptoms. Therefore, we conducted a small pilot study to explore this issue of optimal recall period. Patients were administered sample items from the scale and were then asked a series of questions related to their ability to report their dyspnea using different time periods. Patients rated their dyspnea on a zero to 10 rating scale (zero = lowest possible; 10 = highest possible) with regard to the past 7 days and the past 24 hours. Table 1 indicates that patient ratings of SOB and related FL are highly similar regardless of time frame used. Members of our expert panel decided in light of this that the 7-day time frame would be preferred because it increases exposure to a range of attempted activities, which would produce more complete data per individual.

When asked if they considered the average shortness of breath or their worst shortness of breath in responding to an intensity question, the majority—10 of 14 respondents (71%)—reported using the average for the 7-day recall as well as for the 24-hour recall—9 of 13 respondents (69%). This informed the selection of the following context and recall periods used for dyspnea severity and FL: Dyspnea severity: "Over the past 7 days, how short of breath did you get with each of these activities?"; FL: "Considering your shortness of breath over the past 7 days, rate the amount of difficulty you had when doing the following activities."

Table 1 – Patient ratings of dyspnea and related activity limitation using different recall periods.

	Mean (SD)	Pearson's *r*	*P*
Dyspnea past 7 days	4.5 (2.7)	0.94	< 0.01
Dyspnea past 24 Hours	4.5 (2.8)		
Functional limitation past 7 days	3.9 (2.9)	0.92	< 0.01
Functional limitation past 24 hours	3.9 (2.8)		

SD, standard deviation.

Calibration testing of item pool

Because it is possible that the respondents did not experience all of the activities over a 7-day period, they were allowed to abstain from rating the activities that they did not experience by choosing "I did not do this in the past 7 days." Selecting this response option prompted the respondent to clarify if the reason for not doing the activity had been related to his or her shortness of breath or due to other reasons, including simply not having a chance to do the activity or other health issues. If the reason provided was related to shortness of breath ("I have stopped trying, or knew I could not do this activity because of my shortness of breath"), the response was considered to be equivalent to endorsing "Much difficulty" for the given activity on FL. On the other hand, because either response option does not provide a basis to adequately measure the severity of shortness of breath, neither response choice contributed to dyspnea. As a result, any activities that the respondent has not experienced during the period produced missing responses for dyspnea, regardless of the reason, and for FL if the reason was not related to shortness of breath. For dyspnea, such a scoring approach would likely produce missing response patterns that are most likely nonrandom. That is, the probability of a missing response is increased for people with more dyspnea. For example, patients with severe symptomatic COPD would be more likely to refrain from physically demanding activities and therefore have missing responses on them for dyspnea.

Scaling analysis for dyspnea severity and related FL

Scale construction

A first review of the dyspnea severity and FL items revealed that several activities were not done during a 7-day period for some reason other than dyspnea. Initially, we flagged all items if more than 20% of respondents answered in this manner. Next, we reviewed items to determine if the activity was either more uncommon or a difficult physical task. In total, we identified six items to remove from further consideration as they appeared to be uncommon or unusual and thus would not be good discriminators of the dyspnea experience: sexual activity, light home repair (e.g., fixing a door knob), moderate home repair (e.g., hanging a picture), heavy home repair (e.g., painting), entertaining friends at home, and attending religious services. We also removed certain physical tasks that appeared to be unusually difficult (and thus not done) within this population: vigorous-intensity leisure activity (e.g., football, or tennis); walking (faster than your usual speed) for at least 1 mile (a little more than 1.5 km) without stopping, running or jogging for one-half mile (almost 1 km) without stopping; and running or jogging for at least 1 mile (a little more than 1.5 km) without stopping. We retained the remaining items for the time being, because they all reflected relatively physically challenging activities, which may serve as potential floor items.

Seven additional items (eating, going to the toilet, brushing your teeth, washing your face, visiting friends, working

Table 2 – Dyspnea severity items.

1.	Taking a bath without help
2.	Taking a shower
*3.	Dressing yourself without help
4.	Putting on socks or stockings
5.	Standing for at least 5 minutes.
6.	Walking 10 steps/paces on flat ground at a normal speed without stopping
*7.	Walking 50 steps/paces on flat ground at a normal speed without stopping
8.	Walking mile (almost 1 km) on flat ground at a normal speed without stopping
9.	Walking up 5 stairs without stopping
10.	Walking up 10 stairs (1 flight) without stopping
*11.	Walking up 20 stairs (2 flights) without stopping
12.	Walking up 30 stairs (3 flights) without stopping
*13.	Preparing meals
*14.	Washing dishes
*15.	Sweeping or mopping
16.	Scrubbing the floor or counter
*17.	Making a bed
18.	Lifting something weighing less than 5 lb (about 2 kg, like a houseplant)
19.	Lifting something weighing 5–10 lb (about 2–4.5 kg, like a basket of clothes)
*20.	Lifting something weighing 10–20 lb (about 4.5–9 kg, like a large bag of groceries)
21.	Lifting something weighing more than 20 lb (about 9 kg, like a medium-sized suitcase)
22.	Carrying something weighing less than 5 lb (about 2 kg, like a houseplant) from one room to another
23.	Carrying something weighing 5–10 lb (about 2–4.5 kg, like a basket of clothes) from one room to another
*24.	Carrying something weighing 10–20 lb (about 4.5–9 kg, like a large bag of groceries) from one room to another
25.	Getting in or out of a car
26.	Dining out
27.	Low-intensity leisure activity (gardening, etc.)
28.	Moderate-intensity leisure activity (bicycling on level terrain, etc.)
29.	Walking (faster than your usual speed) for 50 steps without stopping
*30.	Walking (faster than your usual speed) for 1/2 mile (almost 1 km) without stopping
31.	Walking (faster than your usual speed) for at least 1 mile (a little more than 1.5 km) without stopping
32.	Singing or humming
33.	Talking while walking

* Items included in the 10-item clinical trial version of the FACIT-Dyspnea Scale.

at a desk or table, and lying still) were flagged for further examination and subsequently removed because each had a response category with fewer than five cases; such sparse case numbers per category preclude reliable item calibration. The removal of these 17 items reduced the total FACIT-Dyspnea item bank to 33. See Table 2 for all 33 retained dyspnea severity/FL items.

Dimensionality

CFA results using the robust weighted least squares estimator suggested acceptable fit to a unidimensional model (CFI = 0.960; TLI = 0.989). However, the root-mean squared error of approximation was 0.152, indicating less-than-ideal fit. Five pairs of items had residual correlations above 0.20 in absolute value (ranging from 0.204 to 0.262). Reviewing the pairs revealed only one obvious dependency (dressing yourself without help and putting on socks or stockings). EFA results also suggested that a clear and dominant factor is present. The first three eigenvalues were 25.638, 1.402, and 1.083 and the rest were less than 1.0. The ratio of the first to the second eigenvalue was 18.3, and the first factor accounted for about 78% of the total variance, indicating presence of a dominant first factor.

IRT modeling

The graded response model applied to the 33 items fit well, and item calibrations (slope and category threshold locations) enabled us to align the 33 items in each of the two banks along the dyspnea continuum from low score (very healthy) to high score (very impaired).

Scaling missing responses

Once the calibration procedure is complete, one can derive a score for any person who has answered any subset of questions from the calibrated bank. In general, the more questions answered, the more precise the individual estimate. As discussed previously, missing responses on the dyspnea questions may not be ignorable and, hence, treating them as if the questions were not presented could induce bias in the individual estimate. That is, the estimates for patients with more severe COPD symptoms could be systematically lower (albeit small in magnitude) if the responses were treated as missing at random. To further examine the potential effects of nonrandom missing responses, we modeled the missing responses on dyspnea questions using separate IRT modeling. We dichotomized each response into missing (i.e., "I did not do this in the past 7 days.") versus nonmissing (i.e., all of the other responses) and fitted the two-parameter logistic IRT model, which is equivalent to the graded response model with two response categories. We then derived scores based on the 33 dichotomized response variables. The rationale was that if the responses were missing at random, then the modeling of missing responses would produce trait estimates uncorrelated with dyspnea. The correlation between the trait estimates based on the severity of dyspnea and the extent of missing responses was 0.524, indicating that indeed the missing responses contain systematic (scalable) information and hence are not ignorable.

We carried out a similar analysis on the FL questions, where the missing responses were presumed to occur randomly and were therefore unrelated to the trait being measured. Recall that the response was treated as missing only when the response was "I did not do this activity for some other reason (including not having a chance to do it or other health issues)." As expected, the trait estimates derived from modeling the missing responses were practically uncorrelated (r = 0.080) with FL. To further extract scalable information from the dyspnea questions, a similar treatment of missing responses can be applied to the dyspnea answers. That is, if the reason for not doing the activity was related to shortness of breath (i.e., "I have stopped trying, or knew I could not do this activity because of my shortness of breath"), we treat the response as equivalent to indicating "severely short of breath." With this treatment of missing responses, the corre-

lation between the trait estimates based on dyspnea severity and the extent of missing responses became practically zero ($r = 0.076$). The correlation between dyspnea estimates based on the two alternative scoring methods remained very high ($r = 0.997$). Not surprisingly, the correlation between dyspnea and FL scores increased slightly from 0.949 to 0.955. Despite the conceptual distinction, the current dyspnea and FL measures are perhaps too strongly associated with each other to differentiate them empirically. The marginal reliability estimates of the three measures were also very comparable; that is, 0.979, 0.979, and 0.976 for dyspnea, dyspnea with missing handling, and FL, respectively. On the basis of these results, we decided to move forward with the dyspnea scale with missing data handling. We examined the fit statistics [13] of the dyspnea questions. All but one item ("Walking 1/2 mile (almost 1 km) on flat ground at a normal speed without stopping") showed good model fit. Visual inspection of the empirical item characteristic curve of the misfitting item did not reveal any noticeable departures from the theoretical curve. The final graded response model item parameter estimates for dyspnea and FL with missing data handling and CFA factor loadings for FL are available online at: doi:10.1016/j.jval.2010.06.001.

Clinical trial version of the FACIT-Dyspnea scale

We used IRT methodology and expert input to select the most informative and relevant subset of items from among

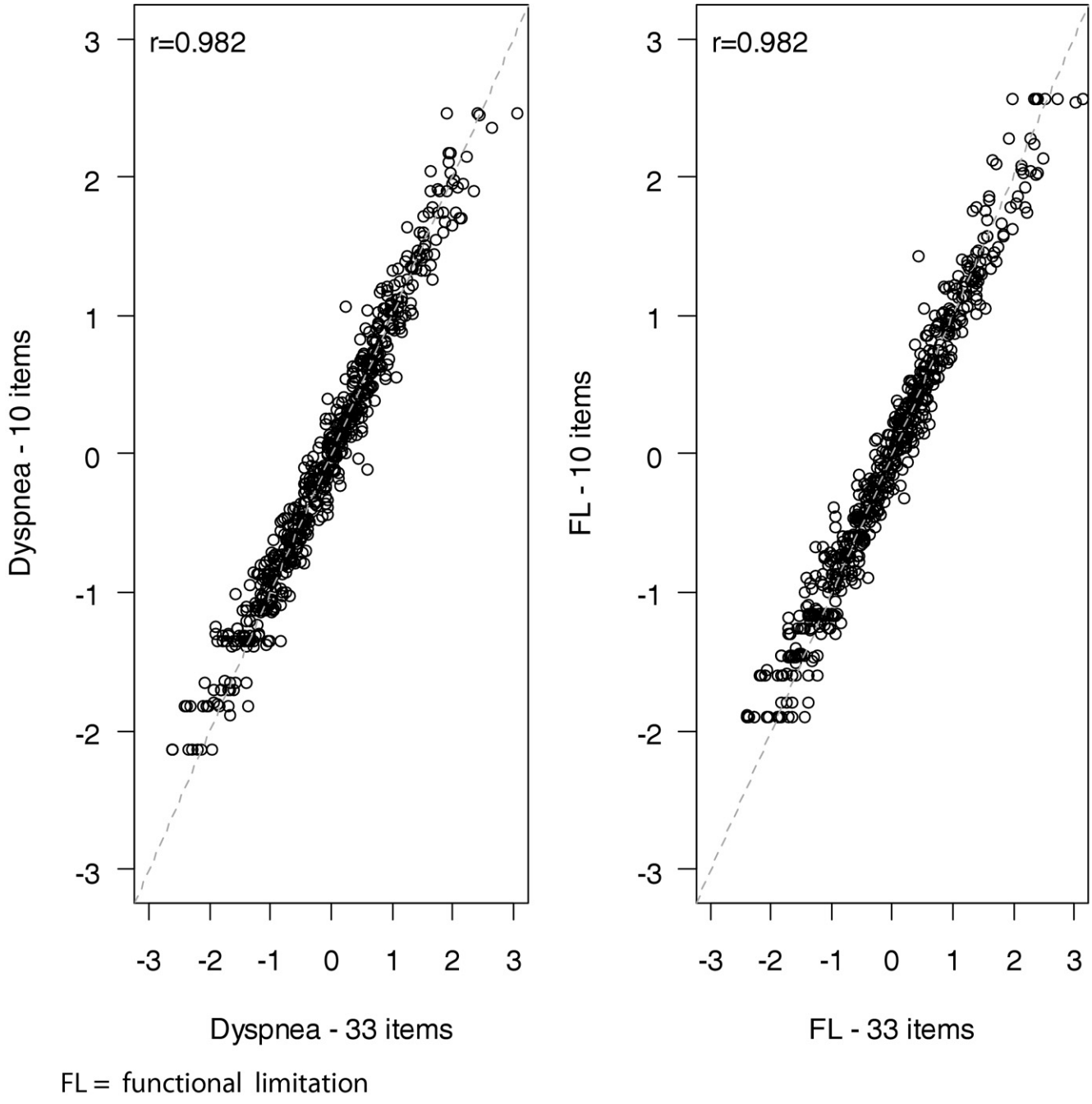


Fig. 2 – Comparison of full-bank and short-form theta estimates.

those in the banks measuring the identified unidimensional concept. We thereby reduced the item bank to a short scale for measuring dyspnea, assuming model fit and sufficient unidimensionality. The reduced scale was constructed in such a way that it retained content validity (i.e., they are true to the subconcepts detailed by patients), and it maximized information gained from patients with COPD. It is critical to recall that with this method of scale construction, questions can be removed and yet comparable scores can be obtained using the retained items. Indeed, it is preferred to exclude one of a set of very highly-correlated items. Doing so provides a more accurate measure of reliability and avoids overemphasis on the narrow sub-concept being tapped by the correlated pair. For example, if two questions about walking are highly correlated, it would be better to use only one, even if one asks about walking 50 steps and the other asks about walking a half-mile. The information gained from one question supersedes the need to ask the other. This is commonplace in a well-constructed bank of questions with multiple response options.

On the basis of the item calibration phase results, we selected 10 items (Table 2) for the clinical trial version of the FACIT-Dyspnea scale. The majority of items were selected because they provided maximum information for the sample. We also considered item local dependence, which reflects excessively high correlation among two or more items, above and beyond what the latent variable predicts. Additional considerations were adequate content coverage across the spectrum of relevant activities and the degree of missing responses in the calibration (scaling) sample.

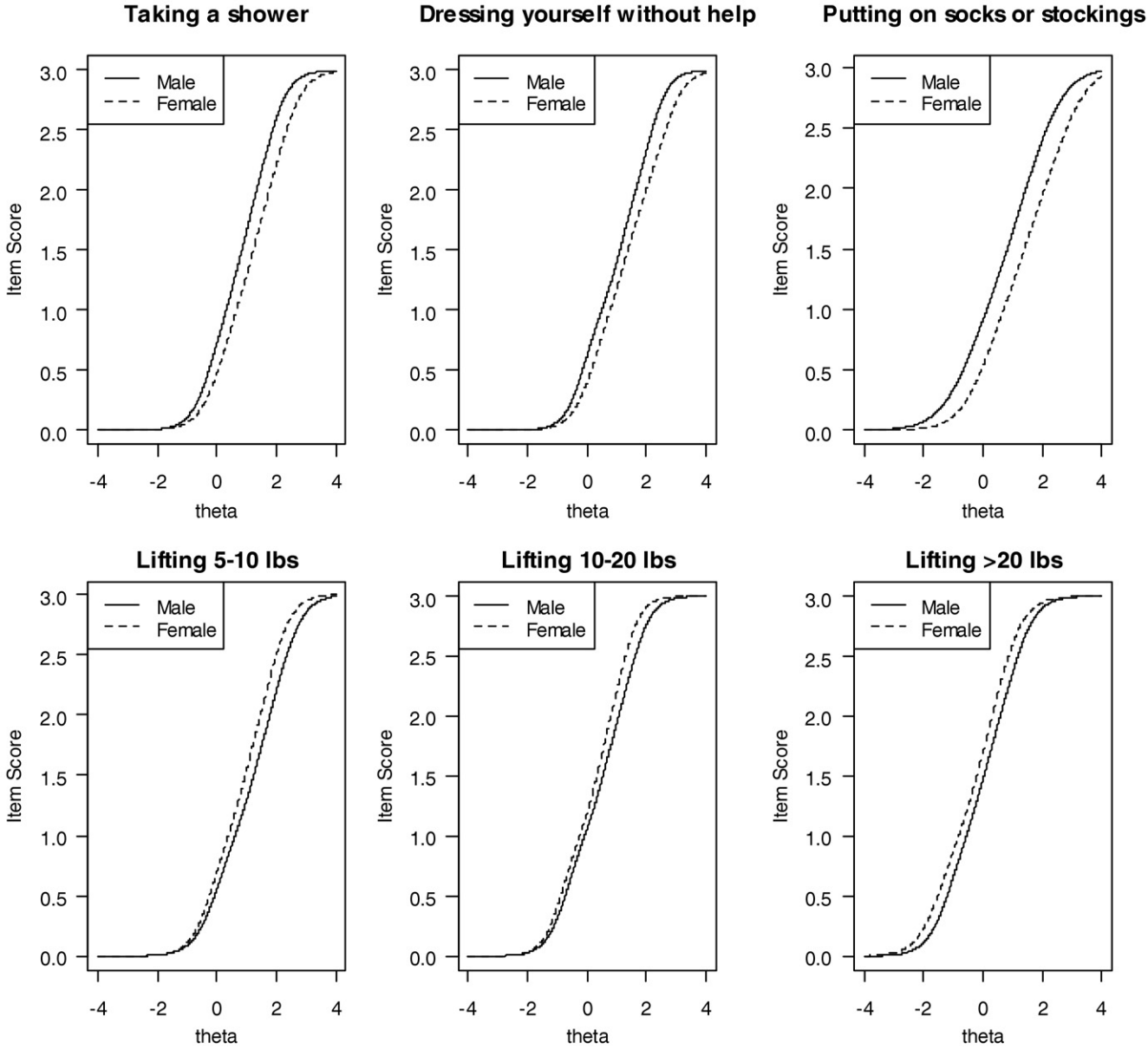


Fig. 3 – Item characteristic curves for differential item functioning (DIF) items.

Scoring of the clinical trial version scale for dyspnea

IRT methodology produces an interval to near-interval score on the trait (θ) being measured. These scores can be estimated with as few as one bank item administered (e.g., using a Bayesian trait estimation procedure such as expected a posterior), but we recommend that at least four to five items be answered to consider the score valid. A unique property of IRT when applied to a well-calibrated item bank is that a common score can be derived even when individual examinees have responded to different sets of items in the pool. Missing item responses are therefore of less concern when applying IRT-based scoring, so dyspnea scores can be directly compared across people answering different subsets of questions from the bank. This is a unique property of IRT, which is made possible by the calibration of all items on a common metric. Figure 2 displays the scatter plot of theta estimates obtained from the 33-item full bank and the 10-item short form. The correlation between the two θ estimates was 0.98. This correlation is about the same as the marginal reliability of the 33-item bank reported earlier. The plotted data included cases with as few as two responses on the 10-item short form.

DIF analysis

A total of six items displayed significant uniform DIF by sex ($P < 0.01$). Three items "favored" women (i.e., less prone to endorse severe response options; for example, severely short of breath) and the other three items favored men. The three items that were easier for males to endorse were taking a shower, dressing yourself without help, and putting on socks or stockings. Conversely, females found the following three items easier to endorse: lifting something weighing 5 to 10 lb, lifting something weighing 10 to 20 lb, and lifting something weighting more than 20 lb. We note that the three items that disfavored men related to self-care, whereas the three items that disfavored women (i.e., prone to report more severe shortness of breath) were related to strength. Interestingly, lifting something less than 5 lb did not show DIF. Figure 3 shows the item characteristic curves (ICC) by sex for the items displaying DIF. The ICCs show that for conditioning, on the dyspnea-severity trait (θ), men and women have different expected scores on the items. When aggregated over items, differences in the ICCs became negligible due to canceling of differences in opposite directions. This implies that, at the scale level, the total expected summed score, as it relates to dyspnea severity, is nearly the same for men and women. The 10-item short form included two items displaying DIF (dressing yourself without help and lifting something weighing 10–20 lb) in opposite directions. To further examine the impact on DIF items on trait estimates, we obtained sex-specific

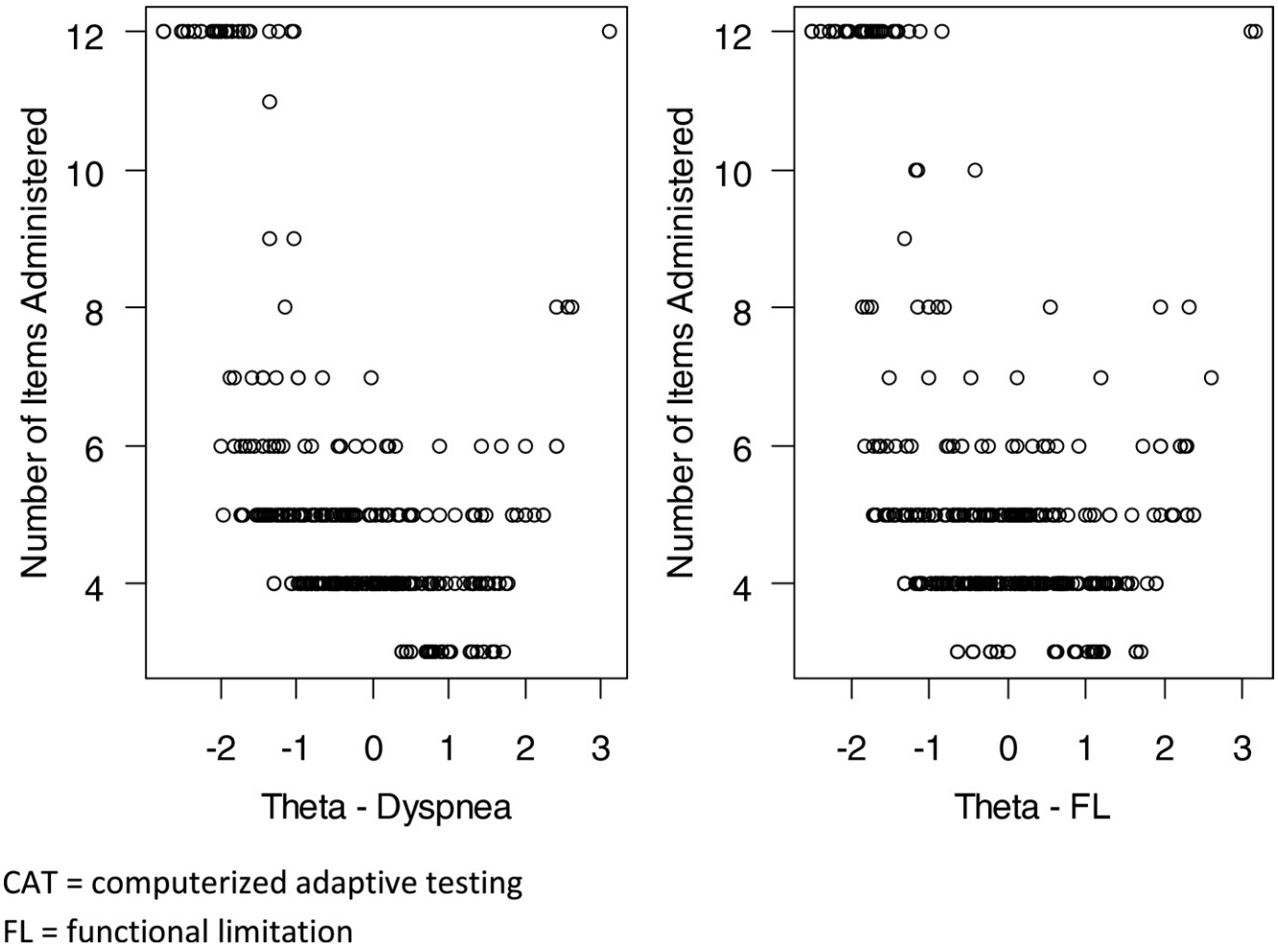
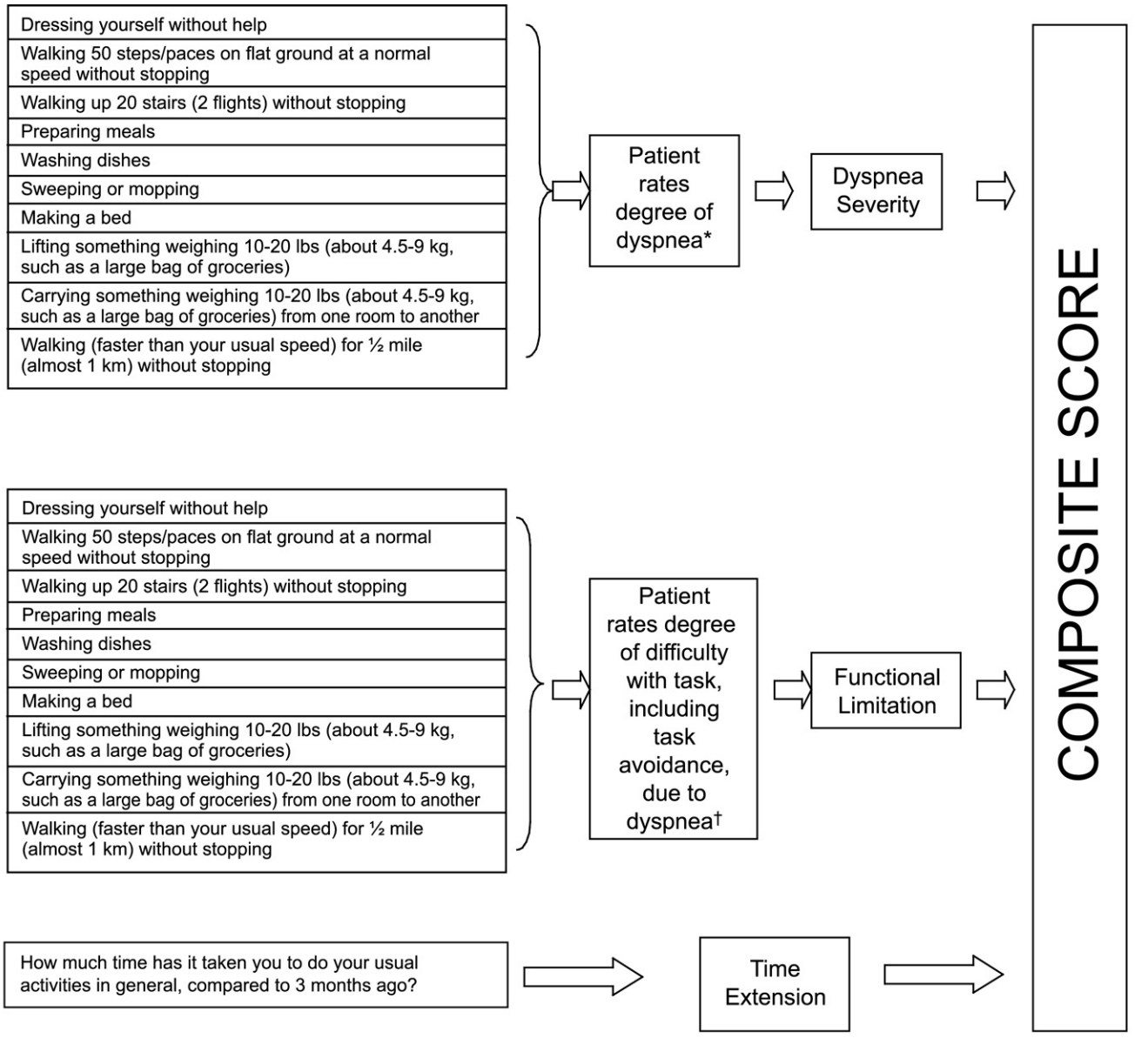


Fig. 4 – Number of items administered under computerized adaptive testing (CAT).

item parameter estimates for the DIF items and compared the DIF-free trait estimates to the initial estimates. The correlation between the two trait estimates was 0.998 and the mean differences by sex were less than 0.01 on the θ metric. This shows that the overall affect of the identified DIF items on trait estimates was negligible at the scale level.

CAT of the Dyspnea bank

The item most frequently selected by CAT was "Carrying something weighing 10-20 lbs (about 4.5–9 kg, similar to a large bag of groceries) from one room to another." On average, CAT administered 4.65 items to the current sample of 608 patients and still maintained a high correlation with the bank theta estimates ($r = 0.97$). About 19% of respondents (n = 115) attained the target SE of 0.3 with only three items, 48% (n = 292) with four items, 19% (n = 117) with five items, and 5% (n = 33) with six items. About 8% (n = 51) were administered seven or more items; 33 (5%) were administered the maximum 12 items. Figure 4 shows the number of items administered by θ estimates. All but one respondent receiving the maximum 12 items were at the bottom (healthy) end of the trait continuum. The maximum number of items to administer can be reduced to eight without loss of precision. Finally, we examined how frequently the items displaying DIF were administered under CAT. The concern was that DIF items under CAT administrations can have a larger impact on trait estimates because their effects may not sum to zero if not balanced (i.e., if items displaying



*Dyspnea Severity Ratings = No shortness of breath, Mildly short of breath, Moderately short of breath, Severely short of breath, I did not do this in the past 7 days.
†Functional Limitations Ratings = No difficulty, A little difficulty, Some difficulty, Much difficulty.

Fig. 5 – Patient-reported outcome conceptual framework of dyspnea severity and functional limitation (10-item calibrated short form).

DIF only in the same direction are selected). The six items flagged for DIF were selected by CAT infrequently (<2%) for the current sample of respondents. None of the six items were among the top seven most discriminating (high-slope) items. Thus, the likelihood of those items being selected under short CATs is low. However, content balancing with a facility to specify a hierarchy of competing (enemy) items can provide a more fundamental solution.

Conceptual frame of dyspnea severity and FL

Figure 5 illustrates the PRO conceptual framework depicting the 20 items measuring dyspnea (10 items) and FL (10 items). The concept of task avoidance is embedded in the FL scale as people are able to respond that they did not do an activity due to shortness of breath. A question pertaining to time extension is included, as it is important to properly interpret change in dyspnea or functional limitations (e.g., if people take longer to do something they may not report an increase in limitation or dyspnea). These three concepts, each measured as indicated, will be combined into a single composite score whereby improvement in one area without decline in another is considered a benefit.

Reliability of the FACIT-Dyspnea bank

Internal consistency reliability

In addition to the marginal reliability (based on IRT trait estimates) reported earlier, we evaluated the internal consistency (coefficient alpha) reliability based on raw item scores. Alpha measures the extent to which the examinees responded consistently across items within a measure or subsets. Cronbach's alpha coefficient for the 33-item dyspnea bank was 0.98; for the 10-item short form it was 0.95. Such high internal consistency estimates indicate that the items included in the dyspnea bank reveal a high level of construct homogeneity [27]. We also computed the item total score correlation as a measure of item quality. The coefficients ranged from 0.64 to 0.87 (mean 0.78 ± 0.06), indicating that the items were very effective in measuring the trait.

Test–retest reliability

This was evaluated by calculating Pearson correlations for the item bank and short form of the dyspnea measure at baseline and 7-day follow-up (range 6 to 12 days; mean 6.5 days). The test-retest reliability coefficients (n = 236) were 0.92 and 0.90 for the 33-item bank and the 10-item short form, respectively. These considerably high test-retest reliability coefficients (i.e., the coefficient of stability) also provide further reassurance as to the stability of these concepts across a week's time, and further support the use of a 7-day recall period. As expected, the 33-item measure showed more stability than the 10-item short form.

Validity of the FACIT-Dyspnea scale

We derived the preliminary validity information based on the scale development sample of 608 people with self-reported COPD. We used several established measures, including the MRC Dyspnea Scale, as external criteria to es-

Table 3 – Correlations among measures.

	Dyspnea-33	Dyspnea-10	Dyspnea-CAT	FL-33	FL-10	FL-CAT	MRC	SF36-PF	SF36-RP
Dyspnea 33	–	0.98	0.97	0.95	0.94	0.93	0.74	−0.86	−0.62
Dyspnea 10	0.98	–	0.97	0.94	0.94	0.91	0.72	−0.85	−0.61
Dyspnea CAT	0.97	0.97	–	0.93	0.92	0.91	0.71	−0.83	−0.61
FL 33	0.95	0.94	0.93	–	0.98	0.98	0.72	−0.89	−0.64
FL 10	0.94	0.94	0.92	0.98	–	0.97	0.70	−0.89	−0.63
FL CAT	0.93	0.91	0.91	0.98	0.97	–	0.70	−0.88	−0.61
MRC	0.74	0.72	0.71	0.72	0.70	0.70	–	−0.69	−0.53
SF36-Physical Function (PF)	−0.86	−0.85	−0.83	−0.89	−0.89	−0.88	−0.69	–	0.70
SF36-Role Physical (RP)	−0.62	−0.61	−0.61	−0.64	−0.63	−0.61	−0.53	0.70	–
SF36-Bodily Pain (BP)	−0.43	−0.42	−0.42	−0.48	−0.47	−0.46	−0.34	0.51	0.50
SF36-General Health (GH)	−0.59	−0.58	−0.57	−0.61	−0.60	−0.59	−0.52	0.61	0.58
SF36-Vitality (VT)	−0.57	−0.56	−0.57	−0.59	−0.57	−0.56	−0.49	0.59	0.64
SF36-Social Function (SF)	−0.56	−0.56	−0.54	−0.60	−0.59	−0.57	−0.46	0.59	0.56
SF36-Role Emotional (RE)	−0.47	−0.46	−0.44	−0.51	−0.51	−0.48	−0.41	0.49	0.66
SF36-Mental Health (MH)	−0.37	−0.36	−0.36	−0.39	−0.39	−0.37	−0.34	0.35	0.41
SF36-Physical Component (PCS)	−0.76	−0.76	−0.75	−0.79	−0.78	−0.77	−0.62	0.89	0.80
SF36-Mental Component (MCS)	−0.36	−0.35	−0.34	−0.39	−0.38	−0.36	−0.33	0.32	0.46
CRQ-SAS Dyspnea	−0.88	−0.87	−0.85	−0.88	−0.87	−0.84	−0.71	0.83	0.67
CRQ-SAS Fatigue	−0.54	−0.53	−0.54	−0.56	−0.55	−0.54	−0.47	0.53	0.59
CRQ-SAS Emotional Function	−0.43	−0.43	−0.43	−0.46	−0.45	−0.44	−0.39	0.41	0.48
CRQ-SAS Mastery	−0.65	−0.63	−0.63	−0.65	−0.64	−0.62	−0.59	0.59	0.55
HADS	0.49	0.47	0.48	0.51	0.49	0.49	0.42	−0.48	−0.52

tablish convergent and divergent validity. Table 3 shows the correlation (validity) coefficients for the three FACIT-Dyspnea measures (i.e., 33-item full bank, 10-item short form, and CAT) with MRC; eight SF-36 subscales, including two component scores; the CRQ-SAS domains (i.e., dyspnea, fatigue, emotional function, and mastery); and the Hospital Anxiety and Depression Scale (HADS). Some measures were scaled in the opposite direction, thus the absolute values of the correlations are most relevant. The three administration formats of the FACIT-Dyspnea scale showed the highest correlations with the CRQ-SAS dyspnea domain (r values between −0.85 and −0.88), followed by the SF-36 Physical Function subscale (r values between −0.83 and −0.86), the SF-36 Physical Component scale (r values between −0.75 and −0.76) and the MRC (r values between 0.71 and 0.74). Because the established MRC has been used to measure dyspnea, the FACIT-Dyspnea scores and the MRC scores were expected to correlate. The strong associations support the concurrent validity of scale scores with a common and well-known dyspnea measure. The correlation between the 2 external dyspnea measures, the MRC and the CRQ-SAS dyspnea domain, was moderate (r = −0.71). The FACIT-Dyspnea measures showed considerably lower correlations with the HADS (r values between 0.47 and 0.49). The CRQ-SAS dyspnea domain also showed strong associations with the SF-36 Physical Function subscale (r = 0.83), and the SF-36 Physical Component scores (r = 0.75), whereas it had a considerably lower association with the SF-36 Mental Component scores (r = −0.45). This pattern of correlations seems to suggest that the CRQ-SAS dyspnea domain is acceptable as a convergent or concurrent criterion for dyspnea measures. In contrast, considerably lower correlations were observed for the three FACIT-Dyspnea measures with the other CRQ-SAS domains; that is, fatigue (r values between −0.53 and −0.54), emotional function (r = −0.43), and mastery (r values between −0.63 and −0.65), providing support for the divergent validity of the measures.

Known groups validity using MRC

We used the MRC as the criterion measure for our known-groups validity analysis. For the calibration sample, we used an adapted version of the scale with a level (zero) appended at the bottom of the classification; that is, zero = no breathlessness; one = I only became breathless after strenuous exercise; two = I only became breathless when hurrying on level ground or walking up a slight hill; three = I had to walk slower than other people on level ground because of breathlessness or I had to stop for breath even when walking at my own pace; four = I had to stop for breath after walking about 100 yards (or after a few minutes) on level ground; and five = I was too breathless to leave the house, or breathless when dressing or undressing. A frequency distribution was calculated on the familiar six-level MRC scale at baseline. Sufficient variability in MRC levels allowed us to conduct cross-sectional analyses of the full item bank and the clinical trial version of the new short-form scores at baseline, using available data from the full item-pool test-

Table 3 (*continued*)

SF36-BP	SF36-GH	SF36-VT	SF36-SF	SF36-RE	SF36-MH	SF36-PCS	SF36-MCS	CRQ-SAS dyspnea	CRQ-SAS fatigue	CRQ-SAS emotional	CRQ-SAS mastery	HADS
−0.43	−0.59	−0.57	−0.56	−0.47	−0.37	−0.76	−0.36	−0.88	−0.54	−0.43	−0.65	0.49
−0.42	−0.58	−0.56	−0.56	−0.46	−0.36	−0.76	−0.35	−0.87	−0.53	−0.43	−0.63	0.47
−0.42	−0.57	−0.57	−0.54	−0.44	−0.36	−0.75	−0.34	−0.85	−0.54	−0.43	−0.63	0.48
−0.48	−0.61	−0.59	−0.60	−0.51	−0.39	−0.79	−0.39	−0.88	−0.56	−0.46	−0.65	0.51
−0.47	−0.60	−0.57	−0.59	−0.51	−0.39	−0.78	−0.38	−0.87	−0.55	−0.45	−0.64	0.49
−0.46	−0.59	−0.56	−0.57	−0.48	−0.37	−0.77	−0.36	−0.84	−0.54	−0.44	−0.62	0.49
−0.34	−0.52	−0.49	−0.46	−0.41	−0.34	−0.62	−0.33	−0.71	−0.47	−0.39	−0.59	0.42
0.51	0.61	0.59	0.59	0.49	0.35	0.89	0.32	0.83	0.53	0.41	0.59	−0.48
0.50	0.58	0.64	0.56	0.66	0.41	0.80	0.46	0.67	0.59	0.48	0.55	−0.52
–	0.47	0.57	0.54	0.45	0.45	0.68	0.43	0.46	0.57	0.48	0.41	−0.50
0.47	–	0.70	0.61	0.51	0.51	0.71	0.54	0.61	0.66	0.57	0.60	−0.59
0.57	0.70	–	0.68	0.58	0.64	0.63	0.71	0.61	0.91	0.72	0.59	−0.71
0.54	0.61	0.68	–	0.59	0.68	0.55	0.77	0.62	0.67	0.70	0.61	−0.71
0.45	0.51	0.58	0.59	–	0.58	0.42	0.78	0.54	0.55	0.59	0.56	−0.62
0.45	0.51	0.64	0.68	0.58	–	0.24	0.91	0.45	0.64	0.89	0.58	−0.86
0.68	0.71	0.63	0.55	0.42	0.24	–	0.23	0.75	0.58	0.35	0.53	−0.41
0.43	0.54	0.71	0.77	0.78	0.91	0.23	–	0.45	0.70	0.86	0.59	−0.84
0.46	0.61	0.61	0.62	0.54	0.45	0.75	0.45	–	0.60	0.52	0.70	−0.55
0.57	0.66	0.91	0.67	0.55	0.64	0.58	0.70	0.60	–	0.75	0.63	−0.73
0.48	0.57	0.72	0.70	0.59	0.89	0.35	0.86	0.52	0.75	–	0.65	−0.88
0.41	0.60	0.59	0.61	0.56	0.58	0.53	0.59	0.70	0.63	0.65	–	−0.66
−0.50	−0.59	−0.71	−0.71	−0.62	−0.86	−0.41	−0.84	−0.55	−0.73	−0.88	−0.66	–

CAT, computerized adaptive testing; FL, functional limitation; MRC, Medical Research Council Dyspnea Scale; SF, short form; CRQ-SAS Dyspnea, Chronic Respiratory Questionnaire–Self-Administered. Standardized/Dyspnea domain; CRQ-SAS Fatigue, Chronic Respiratory Questionnaire–Self-Administered Standardized/Fatigue domain; CRQ-SAS Emotional, Chronic Respiratory Questionnaire–Self-Administered. Standardized/Emotional function domain; CRQ-SAS Mastery, Chronic Respiratory Questionnaire–Self-Administered Standardized/Mastery domain; HADS, Hospital Anxiety and Depression Scale.

ing, focused on differentiating definable ("known") groups defined according to the MRC categories.

In the calibration sample (n = 608), 605 respondents had valid MRC classifications. Using general linear modeling procedures, we confirmed that the MRC levels accounted for the majority of variations in the FACIT-Dyspnea scores ($R^2 = 0.58$, 0.56, and 0.55; $F_{5,599}$=164.2, 150.7, and 143.7, for the 33-item scale, 10-item short form, and CAT, respectively). Subsequent post-hoc pair-wise comparisons revealed statistically significant ($P < 0.05$) mean differences on dyspnea scales (i.e., both 33- and 10-item versions and CAT) across the majority of the MRC categories, supporting the validity of these scale scores (Fig. 6). One exception was that there were no statistically significant mean differences between MRC categories 3 ("I had to walk slower than other people on level ground because of breathlessness or I had to stop for breath even when walking at my own pace") and 4 ("I had to stop for breath after walking about 100 yards [or after a few minutes] on level ground") on all scales. These nonsignificant differences be-

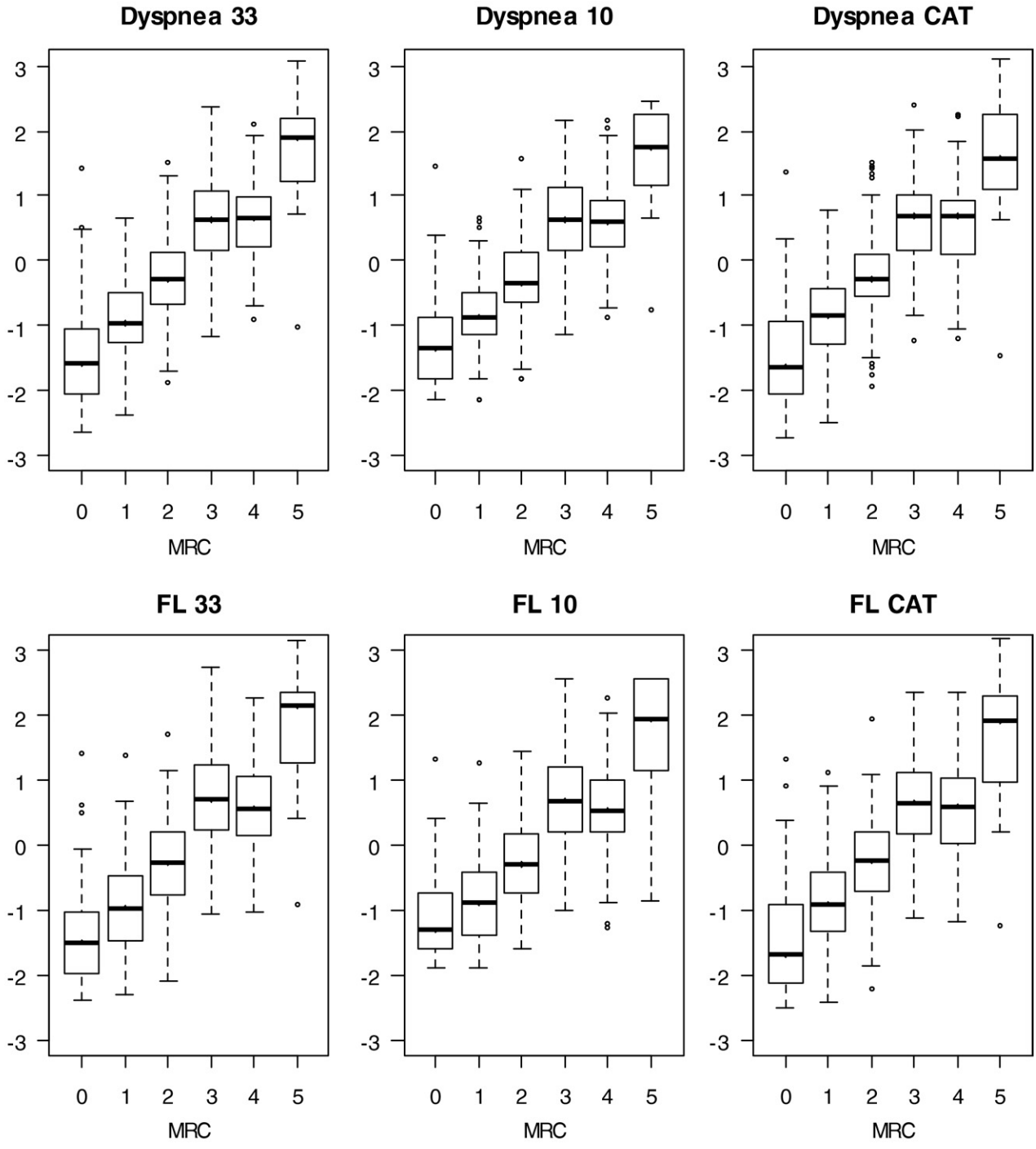


Fig. 6 – Dyspnea/functional limitation score by Medical Research Council dyspnea levels – general population (n = 608).

tween categories 3 and 4 may reflect the semantic similarity of these categories (e.g., both describe needing to stop to catch one's breath). The MRC scale may require further evaluation as a criterion measure to differentiate known groups of dyspnea severity.

Conclusions

Using methods consistent with both the FDA Draft Guidance [4] and the National Institutes of Health Roadmap Patient-Reported Outcomes Measurement Information System initiative [10–12], we developed two item banks to measure dyspnea severity and limitations in function caused by dyspnea, for use in clinical research of people with COPD. These two item banks contain a representative sample of content in each of these two domains, and the item banks satisfy the assumptions and requirements necessary for IRT modeling and item-bank application. As a result, one can select short forms from either of these two banks based upon the presumed severity of dyspnea in the sample to be studied and still express scores on a common metric for dyspnea and related functional limitation. This provides the researcher with unprecedented flexibility in assessment without loss of precision. It also enables use of highly efficient CAT, in which one can anticipate an average of fewer than eight questions per bank in the great majority of cases.

By design, an item bank is overpopulated with questions that represent the concept or "trait" being measured. Therefore, one need not administer all items in the bank to achieve a valid score. Instead, one administers a subset of items from the bank, either predetermined as a static short form or administered in a CAT. Assuming one demonstrates unidimensionality of the concept being measured, and fit of the measurement model as applied to the validation data, an investigator can select items freely from the bank to form a unique scale that measures the concept represented by the bank. Because the items are aligned along a continuum of the trait being measured, such as functional limitation, one could select all items from a restricted area of the continuum (e.g., very demanding or very easy tasks), or across a broad range of function, and in all cases obtain a common functional limitation score for the person answering the subset of questions. Items that are better targeted to the study sample will provide a more accurate score than items that are not well targeted. For example, patients with severe, symptomatic COPD would be less accurately placed with five bank questions asking about climbing several flights of stairs, walking up hills, and carrying out strenuous activities, than if there were five bank questions asking about getting in and out of a car or walking 50 paces. When selecting questions from a bank for a clinical trial, it is advisable to select a range of questions that capture differences between people across the observed continuum of the concept being measured.

This study is not without limitations. It should be noted that the demographic characteristics of participants in our qualitative sample were restricted by geographic, ethnic, and education levels. Future studies may wish to examine the effects of these variables on the experience and self reporting of dyspnea.

As measured by these banks, dyspnea and functional limitation are highly correlated. It is therefore unlikely that independent measurement of both concepts is necessary in any individual study. CAT, or short forms from the banks, such as the 10-item versions presented and tested here, will likely produce reliable and valid estimates of dyspnea or related functional limitations in clinical research projects that include people with COPD. It is important to note that the validity and reliability information presented in this study have been obtained based on cross-sectional data. Further research can demonstrate the clinical validity in longitudinal applications, including responsiveness to change.

Acknowledgments

The authors thank Mary Gabb, MS, a medical writer working with American Healthcare Communications, for editorial input and formatting the article for journal submission.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jval.2010.06.001.

REFERENCES

[1] Mannino DM, Homa DM, Akinbami LJ, et al. Chronic Obstructive Pulmonary Disease Surveillance—United States, 1971-2000. Respir Care 2002;47:1184–99.
[2] Ferrer M, Alonso J, Morera J, et al. Chronic obstructive pulmonary disease stage and health-related quality of life. The Quality of Life of Chronic Obstructive Pulmonary Disease Study Group. Ann Intern Med 1997;127:1072–9.
[3] Edelman NH, Kaplan RM, Buist AS, et al. Chronic Obstructive Pulmonary Disease. Task Force on research and education for the prevention and control of respiratory diseases. Chest 1992;102(Suppl.):243S–56S.
[4] US Department of Health and Human Services. Guidance for Industry. Patient-reported outcome measures: use in medical product development to support labeling claims (DRAFT GUIDANCE). Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071975.pdf [Accessed November 12, 2009].
[5] Chang C-H, Reeve BB. Item Response Theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005;28:264–82.
[6] Cella D, Chang CH. A discussion of item response theory (IRT) and its applications in health status assessment. Med Care 2000;38:1166–72.
[7] Stansfeld SA, Roberts R, Foot SP. Assessing the validity of the SF-36 General Health Survey. Qual Life Res 1997;6:217–24.
[8] Lord FM, Novick M.R. Statistical Theories of Mental Test Scores; Reading, MA: Addison-Wesley, 1968.
[9] Victorson DE, Anton S, Hamilton A, Yount S, Cella D. A conceptual model of the experience of dyspnea and functional limitations in chronic obstructive pulmonary disease. Value Health 2009;12:1018–25.

[10] Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof 2005;28:212–32.
[11] Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap Cooperative Group during its first two years. Med Care 2007;45(Suppl.):S3–11.
[12] DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS Qualitative Item Review. Med Care 2007;45(Suppl.):S12–21.
[13] Orlando M, Thissen D. Further examination of the performance of S-X2, an item fit index for dichotomous item response theory models. Appl Psychol Meas 2003;27:289–98.
[14] Fletcher CM, Elmes PC, Fairbairn AS, Wood CH. The significance of respiratory symptoms and the diagnosis of chronic bronchitis in a working population. Br J Med 1959;2: 257–66.
[15] Samejima F. Estimation of ability using a response pattern of graded scores. Psychometrika Monograph No 17. 1969.
[16] lordif: Logistic Regression Differential Item Functioning using IRT (Version 0.1-9). Available from: http://cran.r-project.org/web/packages/lordif/index.html [Accessed January 11, 2011].
[17] Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and Difwithpar. Med Care 2006; 44(Suppl.):S115–23.
[18] Choi, SW. Firestar: Computerized adaptive testing simulation program for polytomous IRT models. Appl Psychol Meas 2009;33:644–5.
[19] Lareau SC, Carrieri-Kohlman V, Janson-Bjerklie S, Roos PJ. Development and testing of the Pulmonary Functional Status and Dyspnea Questionnaire (PFSDQ). Heart Lung 1994;23:242–50.
[20] Lareau SC, Meek PM, Roos PJ. Development and testing of the modified version of the Pulmonary Functional Status and Dyspnea Questionnaire (PFSDQ-M). Heart Lung 1998;27:159–68.
[21] Leidy NK, Rennard SI, Schmier J, et al. The breathlessness, cough, and sputum scale: the development of empirically based guidelines for interpretation. Chest 2003;124:2182–191.
[22] Garrod R, Bestall JC, Paul EA, et al. Development and validation of a standardized measure of activity of daily living in patients with severe COPD: the London Chest Activity of Daily Living Scale (LCADL). Respir Med 2000;94: 589–96.
[23] Lee L, Friesen M, Lambert IR, Loudon RG. Evaluation of dyspnea during physical and speech activities in patients with pulmonary diseases. Chest 1998;113:625–32.
[24] Eakin EG, Resnikoff PM, Prewitt LM, et al. Validation of a new dyspnea measure: the UCSD Shortness of Breath Questionnaire. University of California, San Diego. Chest 1998;113:619–24.
[25] Schunemann HJ, Griffith L, Jaeschke R, et al. A comparison of the original chronic respiratory questionnaire with a standardized version. Chest 2003;124:1421–9.
[26] Barr JT, Schumacher GE, Freeman S, et al. American translation, modification, and validation of the St. George's Respiratory Questionnaire. Clin Ther 2000;22:1121–45.
[27] Strauss ME, Smith GT. Construct validity: advances in theory and methodology. Ann Rev Clin Psychol 2008; 5:1–25.