Measurement of fatigue: Comparison of the reliability and validity of single-item and short measures to a comprehensive measure

Hee-Ju Kim[a,*], Ivo Abraham[b]

[a] Catholic University of Korea, College of Nursing, Seoul, South Korea
[b] Department of Pharmacy Practice and Science, College of Pharmacy, Department of Family and Community Medicine, College of Medicine, Center for Health Outcomes and Pharmaco-Economic Research, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

ABSTRACT

Background: Evidence is needed on the clinicometric properties of single-item or short measures as alternatives to comprehensive measures.
Objectives: We examined whether two single-item fatigue measures (i.e., Likert scale, numeric rating scale) or a short fatigue measure were comparable to a comprehensive measure in reliability (i.e., internal consistency and test–retest reliability) and validity (i.e., convergent, concurrent, and predictive validity) in Korean young adults.
Methods: For this quantitative study, we selected the Functional Assessment of Chronic Illness Therapy–Fatigue for the comprehensive measure and the Profile of Mood States–Brief, Fatigue subscale for the short measure; and constructed two single-item measures. A total of 368 students from four nursing colleges in South Korea participated. We used Cronbach's alpha and item-total correlation for internal consistency reliability and intraclass correlation coefficient for test–retest reliability. We assessed Pearson's correlation with a comprehensive measure for convergent validity, with perceived stress level and sleep quality for concurrent validity and the receiver operating characteristic curve for predictive validity.
Results: The short measure was comparable to the comprehensive measure in internal consistency reliability (Cronbach's alpha = 0.81 vs. 0.88); test–retest reliability (intraclass correlation coefficient = 0.66 vs. 0.61); convergent validity (r with comprehensive measure = 0.79); concurrent validity (r with perceived stress = 0.55, r with sleep quality = 0.39) and predictive validity (area under curve = 0.88). Single-item measures were not comparable to the comprehensive measure.
Conclusions: A short fatigue measure exhibited similar levels of reliability and validity to the comprehensive measure in Korean young adults.

What is already known about the topic?

- Selection of measures is an important issue in both research and clinical practice.
- Comparative psychometric data for single-item or short fatigue screening measures are limited and further information is needed to enable the potential use of shorter measures.

What this paper adds?

- The five-item Profile of Mood States–Brief fatigue subscale was comparable to the 13-item Functional Assessment of Chronic Illness Therapy–Fatigue in several aspects of reliability and

validity, suggesting that this short fatigue measure can substitute for the comprehensive scale.
- The reliability and validity of single-item fatigue measures (a Likert-type and a numeric rating scale) was much weaker than a 5-item short measure and a 13-item comprehensive measure.

1. Introduction

The selection of instruments to measure psychosocial concepts is an important issue in research and clinical practice (Bowling, 2005; Rosenzveig et al., 2014). Multiple measures exist for a psychosocial concept and may vary in the number of items included and thereby in the level of comprehensiveness. A comprehensive measure comprises multiple items, selected to reflect the domain of a concept as completely as possible.

* Corresponding author at: 505 Banpo Dong, Seo-Cho-Gu, Catholic University of Korea, College of Nursing, Seoul, South Korea.
   E-mail address: heeju0906@gmail.com (H.-J. Kim).

Alternatively, shorter measures may assess the same concept as well, often developed by reducing the number of items in comprehensive measures such as a short measure with five to 10 items, or single-question measures, with either the same or a different scale format compared to the comprehensive instrument. The advantage of a shorter measure is that it imposes less burden on respondents, whether research participants or patients, and can integrate more effectively into research study protocols or clinical visits (Yohannes et al., 2011). In contrast, short and single-item questions may be more vulnerable to inadequate psychometrics and thus impair measurement qualifications (Reissmann et al., 2013).

As part of a research program on short- and single-item measures of psychosocial concepts, we extended our prior work on depression (Kim and Abraham, 2016) to focus on fatigue, similarly in a classical testing-theory framework. Our depression-related study revealed that a selected short measure (i.e., the 5-item Profile of Mood States-Brief [POMS-B] Depression subscale) was comparable to a selected comprehensive measure (i.e., the 20-item Center for Epidemiologic Studies-Depression scale) in validity and reliability, whereas single-item measures were not (Kim and Abraham, 2016). However, the concept of fatigue has a different level of conceptual abstraction and complexity compared to depression, rendering this a unique study.

A single-item global measure may be more appropriate for those concepts that are schematized (Patrician, 2004), unidimensional (Van Hooff et al., 2007), or directly measurable (Reissmann et al., 2013). Indeed, the concept of fatigue has been shown to be unidimensional for healthy participants (Michielsen et al., 2004). Fatigue seems to be unambiguous, and a highly schematized concept for most adults, as it is well summarized in a single question (Van Hooff et al., 2007). Thus, we expected fatigue to be evaluated adequately by shorter measures, whether with a few items or single-item measure.

In parallel to our depression study, the present study on fatigue aimed to examine whether two single-item fatigue measures (i.e., a Likert scale and a numeric-rating scale) or a short fatigue measure are comparable to a comprehensive fatigue measure in reliability (i.e., internal consistency and test–retest reliability) and validity (i.e., convergent, concurrent, and predictive validity) in Korean young adults. Identifying a reliable and valid form of shorter measures is important, because shorter measures are needed when participant boredom, space, time constraints, or cost are major design factors. The present study, too, used a college student sample, a population previously shown to be appropriate for the psychometric evaluation of fatigue measures (Varni and Limbers, 2008). Fatigue is a common complaint in the general population (Yun et al., 2008), including college students (Tanaka et al., 2008).

For measurement comparisons, how researchers conceptualize fatigue or to what degree researchers measure the fatigue experience is important. For instance, the Multidimensional Fatigue Inventory (Smets et al., 1995) defines fatigue in general, mental, and physical dimensions as well as participants' level of desire for activity. Other fatigue measures, such as the Bristol Rheumatoid Arthritis Fatigue Multidimensional Questionnaire, integrates disease-specific dimensions of fatigue: physical, living, cognition, and emotion (Nicklin et al., 2010). Hence, generic versus disease-specific as well as unidimensional versus multidimensional fatigue measures are likely to differ conceptually. We chose to focus on generic fatigue in the general population and reviewed several comprehensive measures as to whether (a) they defined fatigue unidimensionally, (b) measured generic fatigue; (c) had been validated, (d) had been translated into Korean, and (e) included a clinical cut-off point to determine participants with fatigue. We eventually selected the 13-item Functional Assessment

of Chronic Illness Therapy–Fatigue (FACIT-F; Cella et al., 2002) as the comprehensive measure. Selection criteria for the short scale were (a) five to 10 items in composition, (b) a widely used and validated measure, and (c) translated into Korean. We selected the five-item fatigue subscale of the POMS-B (McNair et al., 1992). Lastly, we developed the single-item Likert measure and the single-item numeric-rating scale of fatigue specifically for this study.

2. Method

2.1. Sample and setting

The sample included 368 subjects from four nursing colleges in three metropolitan cities in South Korea (two in Seoul, one in Daegu, one in Pusan) using an instrument set of fatigue, sleep quality, and stress measures. The sole inclusion criterion was that students were taking classes in the selected college; we had no exclusion criteria. We conducted a retest 1 week later at two conveniently selected schools in Seoul. The final sample for the retest was 167. A short retest interval can raise an issue of recall bias; while a longer interval can result in a real change in the measured construct as opposed to evaluating temporal stability. As fatigue status can easily change over time, previous measurement validation studies employed less than a week; for instance, 48 h or less in very unstable disease situations (Paiva et al., 2014) and 1 week in a relatively stable conditions (Takasaki and Treleaven, 2013). Thus, we adopted a retest interval of 1 week, in consideration of the nature of the construct and sample. All participants completed the stress measure, but only 191 completed the sleep-quality questionnaire because it was administered in two conveniently selected schools (one in Seoul and one in Pusan).

2.2. Instruments

2.2.1. Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F)

The FACIT-F (version 4; Cella et al., 2002) is a unidimensional scale with 13-items (Cella et al., 2011). It measures general, not illness-specific, fatigue levels and their impact on daily activities on a 5-point Likert-type scale (0 = not at all, 4 = very much). The FACIT-F has been shown to be valid (e.g., factorial validation, criterion validity) and reliable (0.95 > Cronbach's alphas > 0.90) in the general population as well as in participants with chronic illness (Cella et al., 2011; Smith et al., 2010). The FACIT-F Korean version was validated only linguistically by the test developer. It has been used in various Korean populations (e.g., Yoo et al., 2014). The test developer suggested reversing item scores in calculating the total score, so lower scores indicated higher fatigue. However, in consultation with the test developer (D. Cella, personal communication, 2014), the present study did not reverse the scores to maintain consistency in the directions of scores with other fatigue measures. Total scores in the present study, therefore, ranged from zero to 52, with a higher score indicating more severe fatigue. The test developers identified several useful cut-off points to determine participants with significant levels of fatigue (Cella et al., 2002). Considering our unreversed scores, a score of 35 was two standard deviations above the normal value, defined by population distribution in the United States (sensitivity = 0.34, specificity = 0.97); a score of 30 was higher than the mean of anemic cancer patients (sensitivity = 0.46, specificity = 0.96); and a score of 9 was the cut-off point to best distinguish anemic cancer patients from the general population (sensitivity = 0.92, specificity = 0.69). Because these cut-offs were not based on a clinical fatigue diagnosis, further evidence on validity is needed.