ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 790−801

Journal of Clinical Epidemiology

Multidimensional item response theory models yielded good fit and reliable scores for the Short Form-12 questionnaire

Carlos G. Forero[a,b], Gemma Vilagut[a,b], Nuria D. Adroher[a,b], Jordi Alonso[a,b,*], on behalf of the ESEMeD/MHEDEA Investigators

[a]*Health Services Research Unit, IMIM-Institut Hospital del Mar d'Investigacions Mèdiques, Doctor Aiguader 88, E-08003 Barcelona, Spain*
[b]*CIBER Epidemiología y Salud Pública (CIBERESP), Spain*

Accepted 19 February 2013

Abstract

Objectives: To propose a multidimensional item response theory (MIRT) scoring system for the Short Form 12 (SF-12) with good psychometric properties in terms of fit and reliability.

Study Design and Settings: Two models, indicating physical (PCS) and mental component summary (MCS) dimensions, were fitted to SF-12 data from the European Study of the Epidemiology of Mental Disorders, a representative sample from European adult general population ($n = 21{,}425$; response rate = 61.2%). Goodness of fit, information, reliability, and agreement of individual scores were compared with the classical SF-12 and RAND-12 algorithms.

Results: The bidimensional response process (BRP) model, where all items are indicators of both dimensions, yielded the best fit (root mean square error of approximation = 0.057, comparative fit index = 0.95, and Tucker−Lewis index = 0.94), and highly agreed with PCS and MCS scores from the SF-12 (intraclass correlation coefficients of 0.92 and 0.88, respectively) and RAND-12 (0.88 and 0.95). Regarding reliability, the BRP yielded 0.75 and 0.77 (PCS and MCS, respectively), greater than SF-12 (0.65 and 0.66) and RAND-12 (0.65 and 0.67). As indicated by scale linking, MIRT scores can be interpreted similarly to the classical scores.

Conclusion: The MIRT models showed a clear construct structure for the PCS and MCS dimensions, defined by functional and role limitation content. Results support the use of SF-12 MIRT-based scores as a valid and reliable option to assess health status.

Keywords: Multidimensional graded response model; Confirmatory factor analysis; Health status; Patient-reported outcomes; Outcome measures; Internal consistency

1. Introduction

The Short Form 12 (SF-12) version 1 was developed as a shorter alternative to the SF-36 health survey, for studies in which a 36-item form was too long. Because of its brevity and good performance in clinical assessment, the SF-12 has become a widespread measure of health status and changes in health over time in large samples. It is summarized into two measures, namely the physical (PCS) and mental component summary (MCS) scores [1]. The summaries have successfully been used to detect the presence and severity of physical and mental disorders in clinically defined groups [1,2].

The SF-12 summaries are regression estimates of the corresponding second-order scores in the SF-36 [3−5] and computed as weighted linear composites of individual item responses, coded as dummy variables. Weights are composed of regression coefficients of responses multiplied by the component loading of the item's native SF-36 subscale on its summary [6]. Under the assumption that the items contain physical and mental information, all items participate in the estimation of both components. Another method, the RAND-12 algorithm [7], is arguably the most successful scoring method for the SF-12 based on the item response theory (IRT). It is derived from the application of a Rasch-type IRT model [7,8] to the SF-36 items to obtain eight latent traits. These traits can be summarized in two second-order health scores (physical [PHC] and mental health components [MHC]), originally derived from a two-factor oblique principal axis factor analysis. Like the SF-12, the RAND-12 scores are the product of two regressions of the 36-item PHC and MHC on two six-item

All authors declare that they have neither conflicts of interest nor relevant financial disclosure to inform.

* Corresponding author. Health Services Research Unit, IMIM-Institut Hospital del Mar d'Investigacions Mèdiques, Doctor Aiguader 88, E-08003 Barcelona, Spain, Tel.: +34-933-160-740; fax: +34 399-160-410.
E-mail address: jalonso@imim.es (J. Alonso).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.02.007

What is new?

Key Findings:

- Multidimensional IRT (MIRT) models for scoring the SF-12 questionnaire showed better psychometric properties in terms of fit, construct validity and reliability than more classical score algorithms while being equally interpretable.
- The mental health dimension got especial advantage of the MIRT framework, as it gathered additional information from physical items.

What this adds to what was known?

- MIRT models provide scores for the physical and health dimensions that can be computed independently of the SF36.
- Model based consistency reliability indexes for the different SF-12 score algorithms are provided.

What is the implication and what should change now?

- MIRT models should be used for scoring the SF-12 in applications were measurement precision is a key issue.
- Studies using the SF12 to assess mental health status should used MIRT scores for improved information rather than more classical scoring algorithms.

subsets, weighted by component loadings. Differently from the SF-12, weights are computed from the IRT-weighted items treated as continuous, each item with a single regression weight in its theoretical dimension. The RAND-12 has been shown to be more discriminating than SF-12 in clinical groups [9,10].

Both approaches pose a number of problems, stemming from the fact that, as linear composites aiming at the prediction of the 36-version summaries, the SF-12 and RAND-12 implicitly assume the theoretical and psychometric models of the 36-item versions [4,11,12]. First, score reliabilities depend on a model that is not explicitly stated or estimated. Second, although regression weights optimize prediction of the 36-item summaries, they do not necessarily optimize instrument accuracy. As instrument criterion validity depends on the reliability of the instruments to be correlated, it is a point of major importance [13]. In the case of the SF-12, further difficulties spring from the varying number of alternatives of the 12 items, which violate classical test theory assumptions for computing the alpha coefficient [13−15]. With regard to the RAND-12, the use of a Rasch model (Master's partial credit model) prevents taking full advantage of item information because of the equal-slope restriction [8]. More importantly, the application of regression weights to IRT item weights to predict the RAND-36 PHC and MHC alters the information properties of the IRT weights.

In this article, we aim to provide a model to compute SF-12 scores without having to resort to the prediction of SF-36 summaries. We developed two bidimensional scoring algorithms based on multidimensional IRT graded response models (MGRMs) [16−18], proposing two item structures for the SF-12: items loading in just one dimension and items loading in both dimensions simultaneously. These structures mirror the implicit models of the SF-12 and RAND-12. Scores derived from these structures are compared with those of the original algorithms in terms of reliability.

We hypothesized that using such IRT modeling framework: (1) models with multidimensional response processes at the item level would better capture the properties of the data and yield better fit and more information than unidimensional items, (2) IRT scores would provide individual scores and ordering similar to the standard scoring algorithms, and (3) IRT-based scores would show higher reliability than the scores based on the other algorithms.

2. Methods

2.1. Sample

Data used for this study comes from the European Study of the Epidemiology of Mental Disorders (ESEMeD) [19] project. Briefly, the ESEMeD used a stratified, multistage, clustered area probability sample of noninstitutionalized adult population (aged 18 years or older) of Belgium, France, Germany, Italy, The Netherlands, and Spain. The interviews were conducted between January 2001 and August 2003 using computer-assisted interview techniques. The focus of the study was to estimate the prevalence of mental disorders, using the World Health Organization-Composite International Diagnostic Interview (CIDI) version 3.0. The version 1 of the SF-12 Health Interview (SF-12 v1) was also included to assess the individual general health status. The total sample size achieved was $N = 21{,}425$ individuals, with an overall weighted response rate of 61.2%, ranging from 45.9% in France to 78.6% in Spain.

2.2. Analyses and statistical models

2.2.1. SF-12 v1 and RAND-12 scoring

Version 1 of the SF-12 questionnaire was used. With regard to the SF-12, PCS and MCS scores are short-form estimates of the corresponding SF-36 summaries, computed with the algorithm proposed by its authors [6]: (1) creation of indicator variables (scored 1/0) for item response categories except the one indicating best health state (35 indicator variables), (2) weighted indicator variables were