Journal of Psychosomatic Research 75 (2013) 437–443

Contents lists available at ScienceDirect

Journal of Psychosomatic Research

Adaptive screening for depression — Recalibration of an item bank for the assessment of depression in persons with mental and somatic diseases and evaluation in a simulated computer-adaptive test environment

Thomas Forkmann [a,]*, Ulf Kroehne [b], Markus Wirtz [c], Christine Norra [d], Harald Baumeister [e], Siegfried Gauggel [a], Atilla Halil Elhan [f], Alan Tennant [g], Maren Boecker [a]

[a] *Institute of Medical Psychology and Medical Sociology, University Hospital of RWTH Aachen, Germany*
[b] *Leibniz Institute for Educational Research and Educational Information, Frankfurt/Main, Germany*
[c] *Department of Research Methods, Institute of Psychology, University of Education, Freiburg, Germany*
[d] *Department of Psychiatry, Psychotherapy and Preventive Medicine, Ruhr-University Bochum, LWL University Hospital, Germany*
[e] *Department of Rehabilitation Psychology and Psychotherapy, Institute of Psychology, Albert-Ludwigs-University, Freiburg, Germany*
[f] *Department of Biostatistics, Faculty of Medicine, Ankara University, Turkey*
[g] *Department of Rehabilitation Medicine, Faculty of Medicine and Health, University of Leeds, UK*

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form 27 August 2013
Accepted 28 August 2013

Keywords:
Computer-adaptive test
CAT
Depression
Item response theory
Rasch model
Screening

ABSTRACT

Objective: This study conducted a simulation study for computer-adaptive testing based on the Aachen Depression Item Bank (ADIB), which was developed for the assessment of depression in persons with somatic diseases. Prior to computer-adaptive test simulation, the ADIB was newly calibrated.
Methods: Recalibration was performed in a sample of 161 patients treated for a depressive syndrome, 103 patients from cardiology, and 103 patients from otorhinolaryngology (mean age 44.1, SD = 14.0; 44.7% female) and was cross-validated in a sample of 117 patients undergoing rehabilitation for cardiac diseases (mean age 58.4, SD = 10.5; 24.8% women). Unidimensionality of the itembank was checked and a Rasch analysis was performed that evaluated local dependency (LD), differential item functioning (DIF), item fit and reliability. CAT-simulation was conducted with the total sample and additional simulated data.
Results: Recalibration resulted in a strictly unidimensional item bank with 36 items, showing good Rasch model fit (item fit residuals < |2.5|) and no DIF or LD. CAT simulation revealed that 13 items on average were necessary to estimate depression in the range of −2 and +2 logits when terminating at SE ≤ 0.32 and 4 items if using SE ≤ 0.50. Receiver Operating Characteristics analysis showed that θ estimates based on the CAT algorithm have good criterion validity with regard to depression diagnoses (Area Under the Curve ≥ .78 for all cut-off criteria).
Conclusion: The recalibration of the ADIB succeeded and the simulation studies conducted suggest that it has good screening performance in the samples investigated and that it may reasonably add to the improvement of depression assessment.

Introduction

Self-report instruments, also referred to as patient-reported outcomes (PRO), are a common means of identifying depression in routine clinical practice and research. Many such questionnaires have been developed and persuasive psychometric characteristics have been reported for these instruments based upon Classical Test Theory (CTT) assumptions [1,2].

However, in the last years it was demonstrated that PRO could benefit substantially from modern approaches such as item response theory (IRT) [3]. Generally, applying IRT models can provide additional perspectives on instruments used for depression diagnostics, such as revelation of item bias across subgroups [4–7], infringement of unidimensionality [8,9], or redundancies in the item sets [10]. Consequently, some potential for further improvement of depression specific PROs is evident. Because of its particular desirable properties such as parsimony and similar differentiation of items the one-parameter Rasch model, a member of the group of IRT models, was used for the present study [11,12].

A recent and probably most appealing new perspective offered by IRT is the implementation of Computer-Adaptive Testing (CAT). CAT chooses and presents targeted items from a calibrated item bank to the respondent, thereby minimizing the standard error of measurement

* Corresponding author.
E-mail addresses: tforkmann@ukaachen.de (T. Forkmann), kroehne@dipf.de (U. Kroehne), markus.wirtz@ph-freiburg.de (M. Wirtz), Christine.Norra@rub.de (C. Norra), harald.baumeister@psychologie.uni-freiburg.de (H. Baumeister), sgauggel@ukaachen.de (S. Gauggel), ahelhan@yahoo.com (A.H. Elhan), a.tennant@leeds.ac.uk (A. Tennant), mboecker@ukaachen.de (M. Boecker).

0022-3999/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jpsychores.2013.08.022

(SEM) and reducing test length [13,14]. Simulation studies demonstrate that CAT may measure sufficiently precise with approximately six items [15,16].

The central foundation stone of each unidimensional CAT is a calibrated item bank [17]. This is a set of items with proven unidimensionality for measuring the latent variable and with item difficulties capturing a wide range of this dimension. Items are calibrated, i.e., estimates of item parameters (such as the item difficulty) are provided for each item.

Forkmann and colleagues [18] developed the Aachen Depression Item Bank (ADIB) that has been calibrated on a mixed sample of both persons with primarily mental illnesses (depression) and primarily somatic illnesses (persons with cardiac or otorhinolaryngologic diseases). Using the software WINSTEPS 3.60.1 Forkmann et al. [18] showed that the ADIB is essentially unidimensional, fits the Rasch model and captures a wide range of the latent continuum. The ADIB proved to be useful for the derivation of high quality static short scales for the assessment of depression supporting its general psychometric quality [19–23]. However, Forkmann et al. [18] further reported that there were small signs of a potential secondary dimension constituted by items about suicidal ideation and behavior. This finding might be interpretable in line with the assumption that suicidal ideation and behavior might have to be considered as a nosological entity itself [24]. Signs for multidimensionality were only minor. Nevertheless, strict – as opposed to essential – unidimensionality is necessary for bias-free estimates in CAT procedures which requires a more rigorous statistical approach [[25,26], and methods section]. Furthermore, local independence was assessed using a less rigid criterion than necessary if the item bank should be used for computerized adaptive testing so that the recalibration reported in the present study appeared inevitable.

The current study had three aims. The first aim was to conduct a recalibration of the ADIB through secondary analysis of data from the study of Forkmann et al. [18] using more strict criteria in order to improve unidimensionality, local independence and reduce DIF. Based on a thoroughly calibrated item bank a CAT program could be build, which is the final aim of the item bank development. A CAT that accesses an item bank calibrated on patient samples with mental and somatic diseases would help to reduce time and test burden, enhance precision of measurement and allow for bias free estimates of depression severity independent of somatic diseases. Based on more economic, precise and bias-free depression measurements it is conceivable that therapeutic interventions could be targeted more purposefully to the patient.

The second aim was to cross-validate the new calibration of the ADIB on an independently drawn sample of patients undergoing rehabilitation for cardiac diseases. The third aim was to conduct a preliminary simulation study in order to test the item bank's performance in a simulated CAT environment with regard to its precision, economy, and the validity of the interpretation of θ estimates based on the CAT. In a *real* CAT each patient fills in adaptively presented items at the computer. By contrast, in a *simulated* CAT, paper and pencil data on the items of the bank are treated as if they had been collected adaptively. That means that the algorithm chooses a first item of medium difficulty and then, based on the real answer given by the patient, the next item is chosen. Before real CAT application, CAT is usually used in simulation studies to see whether further improvement is necessary.

Methods

Samples

The recalibration of the item bank (step I) was conducted through a secondary analysis of data reported in Forkmann et al. [18] that was recruited from a German university hospital and a community psychiatric clinic (sample I; N = 367: 161 patients treated for a depressive syndrome (DP), 103 patients from cardiology (CP), and 103 patients from otorhinolaryngology (OP)). Participants' average age was 44.1 (SD = 14.0) and 44.7% were female (see [18] for details).

Cross-validation (step II) was conducted on a newly drawn sample of persons undergoing rehabilitation for cardiac diseases (sample II; N = 117, M_{age} 58.4, SD = 10.5; 24.8% women). Participants suffered from ischemic heart disease (ICD-10: I20-I25; 62.4%), other forms of heart disease (I30-I52; 13.7%), both (13.7%), or essential primary hypertension (I10; 9.4%). Recalibration was based on sample I because widespread depression severity levels across the latent continuum and a balanced composition of the sample in terms of patients with a mental disorder and patients with a primary somatic disease are good for stable item calibrations.

The CAT simulation (step III) was performed based on two sources of data: (1) real patient data of samples I (N = 367) and II (N = 117) and (2) a sample of N = 500 normally distributed simulees (S_s).

All participants took part voluntarily without payment and signed an informed consent prior to testing. General inclusion criteria were German language skills and the ability to concentrate for 1.5 h. Additionally, participants from the psychiatric clinic were consecutively included if they reported mood disorders at admission. Test administration was conducted by trained personnel. The study was approved by the local ethics committee and performed according to the Declaration of Helsinki [27].

Material

a) Aachen Depression Item Bank (ADIB)

The ADIB consists of 79 items referring to cognitive/emotional aspects of depression [18]. All items are scaled using a 5-point Likert scale with the response categories "*never*" (0) "*rarely*" (1), "*sometimes*" (2), "*mostly*" (3), and "*always*" (4). The introducing phrase for each item is "How often during the last two weeks…" [see [18] for details].

b) Demographic data sheet

Participants filled in a demographic data sheet. Clinical data were taken from medical records.

Data analysis

Analyses consisted of three steps and are described in detail in the following paragraphs.

Step I) Initial evaluation of unidimensionality of the ADIB

Dimensionality of the ADIB was initially evaluated applying exploratory factor analysis (EFA) for categorical data using weighted least square methods and PROMAX rotation in the program MPlus [28]. We used the root mean error of approximation (RMSEA), which values model parsimony for determination of model fit. RMSEA values <0.08 indicate sufficient fit. Items were assigned to a specific factor if factor loadings were ≥ 0.4. Items with cross-loadings, i.e., factor loadings ≥ 0.4 on more than one factor, were removed [29]. RMSEA can also be used to determine the appropriate number of factors by specifying a series of models of increasing complexity. Then, the model that fits the data well, and that fits the data better than any other model with more or fewer factors is chosen [30].

Step II) Evaluation of Rasch model assumptions, recalibration and cross-validation of the item bank

In this step, a set of Rasch model assumptions was tested. All analyses were performed with the program RUMM 2030 [31] using the Partial Credit Model [PCM; 32]. Details on the following steps of analyses can be found elsewhere [12,33].

(1) *Ascending ordering of response categories*. For each item, it was examined if response categories were adequately ordered. In the case of disordered thresholds adjacent categories can be merged, which usually results in better fit of the model [12,33].