International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 18(1): 23–36 (2009)
Published online 4 February 2009 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.274

Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application

HERBERT FLIEGE,[1] JANINE BECKER,[1] OTTO B. WALTER,[2] MATTHIAS ROSE,[3] JAKOB B. BJORNER[3] & BURGHARD F. KLAPP[1]

1 Clinic for Internal Medicine, Department of Psychosomatic Medicine and Psychotherapy, Charité Universitätsmedizin Berlin, Germany
2 Institute of Psychology, University of Münster, Germany
3 QualityMetric Incorporated, Lincoln, RI, USA

Key words
depression, diagnostic assessment, questionnaire, computer-adaptive testing, item response theory

Correspondence
Herbert Fliege, Department of Psychosomatic Medicine and Psychotherapy, Charité Universitätsmedizin Berlin, Luisenstrasse 13 A, D-10117 Berlin, Germany.
Telephone (+49) 30-450-553 097
Fax (+49) 30-450-553 989
Email: herbert.fliege@charite.de

Received 22 June 2007;
revised 29 November 2007;
accepted 31 January 2008

Abstract

In the past, a German Computerized Adaptive Test, based on Item Response Theory (IRT), was developed for purposes of assessing the construct depression [Computer-adaptive test for depression (D-CAT)]. This study aims at testing the feasibility and validity of the real computer-adaptive application.

The D-CAT, supplied by a bank of 64 items, was administered on personal digital assistants (PDAs) to 423 consecutive patients suffering from psychosomatic and other medical conditions (78 with depression). Items were adaptively administered until a predetermined reliability ($r \geq 0.90$) was attained. For validation purposes, the Hospital Anxiety and Depression Scale (HADS), the Centre for Epidemiological Studies Depression (CES-D) scale, and the Beck Depression Inventory (BDI) were administered. Another sample of 114 patients was evaluated using standardized diagnostic interviews [Composite International Diagnostic Interview (CIDI)].

The D-CAT was quickly completed (mean 74 seconds), well accepted by the patients and reliable after an average administration of only six items. In 95% of the cases, 10 items or less were needed for a reliable score estimate. Correlations between the D-CAT and the HADS, CES-D, and BDI ranged between $r = 0.68$ and $r = 0.77$. The D-CAT distinguished between diagnostic groups as well as established questionnaires do.

The D-CAT proved an efficient, well accepted and reliable tool. Discriminative power was comparable to other depression measures, whereby the CAT is shorter and more precise. Item usage raises questions of balancing the item selection for content in the future.

Introduction

Depression is one of the most prevalent and relevant psychopathological phenomena in various clinical settings (Attkisson and Zich, 1990; Katon and Sullivan, 1990). The World Mental Health Surveys show that there is a strong association between depression and/or anxiety and physical conditions (Scott *et al.*, 2007). Very likely,

depression is not only a frequent consequence of physical illness but also a predictor for various physical outcomes (Gaynes *et al.*, 2002; Katon and Ciechanowski, 2002), for example pain development (Chou, 2007), heart failure mortality (de Denus *et al.*, 2004) or breast cancer recurrence (Groenvold *et al.*, 2007).

Clinicians face a major task in recognizing, diagnosing and treating depression. A precise and economic measurement of the level of depression is therefore fundamental. A number of established questionnaires based on principles of classical test theory (CTT) meet these requirements (Gilbody *et al.*, 2001), e.g. the Hospital Anxiety and Depression Scale (HADS: Zigmond and Snaith, 1983), the Beck Depression Inventory (BDI: Beck and Steer, 2003), the Patient Health Questionnaire (PHQ-9: Kroenke *et al.*, 2001) or the Zung Self-Rating Depression Scale (Zung, 1965). However, CTT based instruments have certain drawbacks, the most prominent one being a low measurement precision for extreme (low or high) levels of the latent trait. To overcome this problem within the framework of CTT, an additional number of items were needed.

More recently, health assessment researchers have been increasingly interested in applying Item Response Theory (IRT) methods to questionnaire development, evaluation and refinement (Edelen and Reeve, 2007). While CTT assumes that the same measurement precision applies throughout the measurement range, IRT allows for the assessment of measurement precision for each level of the measured construct. Thus, IRT methods can identify the range of the latent trait continuum for which the item can best discriminate among individuals, and reveal how well different items discriminate at particular levels of the construct. This information can be used to select the most informative (discriminative) items and to administer only them to the test taker.

In some studies, the IRT approach was applied to investigate and improve established CTT based questionnaires for depression. In all of these studies, IRT methods disclosed item information that could be used to improve the measurement properties of the fixed-form questionnaires (Baer *et al.*, 2000; Bech *et al.*, 2001; Edelen and Reeve, 2007; Kim *et al.*, 2002; Meijer and Baneke, 2004; Olsen *et al.*, 2003; Orlando *et al.*, 2000; Stansbury *et al.*, 2006).

Computer-adaptive testing (CAT) algorithms based on IRT also offer attractive opportunities for simultaneously optimizing both measurement precision and economy (Walter *et al.*, 2007). A CAT algorithm uses information from questions already answered in order to select the most appropriate question to be administered next. By asking the most appropriate questions for each individual, it becomes possible to administer fewer items and yet achieve greater measurement precision across the entire range of a construct like depression. This also reduces floor and ceiling effects (Embretson and Reise, 2000). Although the theoretical advantages of CAT are widely recognized, up until now there have been only few reports of pioneer CAT applications measuring health constructs in clinical settings (Handel *et al.*, 1999; Walter *et al.*, 2007; Ware *et al.*, 2003).

In an earlier study, we endeavoured to develop pilot versions of a CAT for depression (Fliege *et al.*, 2005) and for anxiety (Walter *et al.*, 2005), based on IRT. All instruments were in German. The Depression-CAT (D-CAT) was designed to measure the level of depression. Initially, we started with 144 items that originated from 10 fixed-form questionnaires and were indicative of depression according to the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV) criteria (American Psychiatric Association, 1994). After testing for unidimensionality of the items and item characteristics, 64 items remained in the pilot CAT item pool.

The properties of the item pool were pre-evaluated on simulated data varying in the level of depression and on real patients' data comprising of all items. In simulation studies, the D-CAT proved to be economic and precise. Correlations between CAT scores and theta scores estimated from all items were high ($r = 0.95$). Correlations between the CAT and classical questionnaires for depression, the BDI ($r = 0.79$) and the Centre for Epidemiological Studies Depression (CES-D) scale ($r = 0.85$), were satisfactory.

Thus, the pilot version of the D-CAT proved to be reliable, valid and economical. However, it had yet to be evaluated prospectively.

In the present study, we applied the D-CAT algorithm to patients, with the intention of evaluating and validating the pilot version of the tool under real clinical conditions.

An additional aim was to investigate patients' acceptance of the computerized test administration. Computer-administered testing has become increasingly popular and studies suggest that the technology is basically accepted by patients (Allenby *et al.*, 2002; Bendtsen and Timpka, 1999; Carlson *et al.*, 2001; Kobak *et al.*, 1996; Rose *et al.*, 2002). However, this has yet to be proven with respect to computerized adaptive testing, where item wordings and response formats may also differ in the course of test taking.

Subjects and methods

Sampling and assessment design

The D-CAT was administered to a convenience sample of consecutive inpatients treated between July and October 2005 at the Department of Psychosomatic Medicine and Psychotherapy, Charité Universitätsmedizin Berlin, Germany. Patients were included if they stayed in hospital for at least one week. In 5% of the cases no psychological testing was carried out due to insufficient language mastery, in another 3% due to current physical conditions. All other patients were willing and able to participate. The sample totals 537 patients. Patients were administered the D-CAT along with several other instruments assessing depression. While all patients completed the D-CAT, the various validation instruments were not administered to all patients, in order to keep patient burden low. The total sample was separated into sample A with 423 patients and sample B with 114 patients.

Sample A served to investigate feasibility, patients' acceptance, and item usage of the CAT. It was also used to investigate convergent validity. For this purpose, we administered the CES-D scale and the HADS to a subsample of 127 patients (A1) and the BDI to another subsample of 111 patients (A2) (instrument description see later). The remaining patients of sample A completed several questionnaires assessing anxiety as part of a different study reported elsewhere.

Samples A and B were used to investigate the discriminative validity of the CAT. In sample A, clinical diagnoses were used. In sample B, standardized diagnoses were used to investigate the discrimination between patients diagnosed with a depression versus patients diagnosed with a mental or behavioral disorder other than depression. Clinical diagnoses were given by an experienced physician and/or psychologist at the end of treatment, involving a mean length of 19 hospital days. Diagnoses were based on the gathered clinical information and supported by software for coding diagnoses (Diacos®). Sample B was additionally diagnosed by means of fully structured face-to-face interviews using the computerized version of the Composite International Diagnostic Interview (CIDI) [Wittchen and Pfister, 1995; World Health Organisation (WHO), 1997]. All interviewers had received appropriate, certified training (Prof. Wittchen, University of Dresden, Germany). For comparing discriminative validity with established questionnaires, sample B completed the D-CAT along with the HADS and the BDI.

Questionnaire assessment took place within the first two days of inpatient treatment. Standardized diagnostic interviews took place within the first week of inpatient stay.

Sample characteristics

The overall patient sample ($N = 537$) was comprised of 377 women (70.2%) and 160 men (29.8%). The mean age was 42.0 years [standard deviation (SD) = 15.1, range 18–77 years]. Women were slightly overrepresented in sample B (78.1%) compared to sample A (68.1%). There were no age differences between any of the sub-samples.

The main clinical diagnoses according to the International Classification of Diseases-10th Revision (ICD-10) F were (in order of prevalence) somatoform disorders (F45; $n = 129$; 25.3%), depressive disorders (F3; $n = 124$; 24.4%), adjustment disorders/stress reactions (F43; $n = 94$; 18.5%), eating disorders (F50; $n = 81$; 15.9%), anxiety disorders (F40/41; $n = 49$; 9.6%), dissociative [conversion] disorders (F44; $n = 14$; 2.8%), substance abuse/addiction (F1; $n = 10$; 2.0%) and other disorders (F6, F0, F42; $n = 8$; 1.5%). Twenty-eight patients suffered from other medical conditions but were not diagnosed with an ICD-10 F-diagnosis. The diagnostic distribution did not differ between sub-samples A1 and A2. However, subsamples A and B differed in that depressive disorders were more frequently diagnosed in sample B ($n = 46$; 40.3%) than in sample A ($n = 78$; 18.4%), whereas adjustment disorders were more frequently diagnosed in sample A ($n = 86$; 22.2%) than in sample B ($n = 8$; 7.0%).

According to the CIDI interviews, 55 of the patients from sample B (48.2 %) were diagnosed with depression, whereas 59 of the sample B patients (51.8 %) did not fulfill the diagnostic criteria for depression.

Instruments

Computer-adaptive test for depression (D-CAT)

The D-CAT makes use of a pool of 64 items. In the initial item bank development, 144 items were examined for unidimensionality, local independence and item characteristics (response curves, slope parameters, reliability curves, test information curves, differential item functioning, threshold, and location parameters) (Bjorner *et al.*, 2003; Ware *et al.*, 2000). The subset of 64 items defined a sufficiently unidimensional item bank and had good item properties. These items cover nine groups of depression symptoms along the lines of the DSM-IV criteria (American Psychiatric Association, 1994): (1) 25 items address aspects of depressed mood, ranging from extreme levels like 'not able to cheer up' or 'unbearably sad' to polar opposites like 'feeling happy' or 'enjoying

life' (e.g. 'During the last week I felt depressed'). Associated with depressed mood are feelings of tension, anxiety or insecurity (eight items) and depersonalisation/derealisation (one item), as listed in the DSM-IV under the first diagnostic criterion. Other items address (2) activity disturbance (seven items, e.g. 'I couldn't bring myself to work'), (3) fatigue or loss of energy (seven items, e.g. 'I feel easily fatigued'), (4) self-reproach or feelings of guilt (seven items, e.g. 'I feel self-confident'/scoring reversed), (5) loss of interest and pleasure (three items, e.g. 'I lost interest in other people'), (6) poor concentration or indecisiveness (three items, e.g. 'I had difficulties concentrating'), (7) thoughts of death or suicide (one item, 'I had thoughts of taking my own life'), (8) sleep disturbance (one item, 'I wished I could sleep long and deeply') and (9) appetite or weight loss (one item, 'my appetite is diminished').

The original item wordings and response formats were not altered by us in any way, because CATs allow for combining items that vary in format. Thus, 23 items refer to the momentary state. For 26 items the recall period was one week. Seven items refer to four weeks and for eight items no recall period is specified.

To estimate the latent trait, or theta, we used the Generalized Partial Credit Model (GPCM) by Muraki (1992), which is a two parameter model allowing items to have different slopes. The CAT algorithm was developed and programmed by Dr Otto Walter as in-house software (Walter *et al.*, 2005, 2007). The development of the D-CAT was described elsewhere in more detail (Fliege *et al.*, 2005).

The CAT algorithm starts with the item that has the highest level of information regarding the theoretical mean latent trait score of zero ('During the past week I felt depressed') and that is at the same time a good indicator of the latent trait construct. Item information has been determined in earlier studies (Fliege *et al.*, 2005). The algorithm uses the subject's response to this item to estimate the latent trait using the expected a posteriori method (EAP) (Bock and Mislevy, 1982). Subsequently, the algorithm selects further items based on the highest possible information for the current latent trait score. The latent trait is estimated after each item administration on the basis of the accumulated information together with the information from the new response. Measurement precision is calculated as a confidence interval after each response. The adaptive testing stops when a predefined measurement precision is attained. We set the measurement precision at a standard error (SE) of ≤ 0.32, which corresponds to a reliability of $r = 0.90$. With this stopping rule, the complete test cycle has a variable length, depending on the individual responses and the point at which the stopping rule value is attained. We did not preset a minimum number of item administrations. Thus, theoretically, the algorithm could stop after an administration of only two or three items, given that the item information is high enough for the score estimation to reach the predefined minimum reliability.

In the field of IRT research, it is common practice to report theta scores on a metric with a mean of zero and a SD = ±1 (Embretson and Reise, 2000). In our earlier papers, we kept to this metric when reporting scale development and item characteristics. However, with respect to clinical application, we wanted to avoid negative values and potential problems with their interpretation. With a view to a metric that is more easily interpretable, we followed other examples of CAT applications (Ware *et al.*, 2003) and opted for using *T*-values with a mean of 50 and a SD of 10. High values indicate a higher level of depression and vice versa.

Validation instruments

To validate the real CAT application, we included three static, CTT-based questionnaires that are used extensively for measuring depression in medical hospital patients, in primary care, and in a broad range of research settings.

The HADS (Herrmann *et al.*, 1995; Zigmond and Snaith, 1983) is a 14-item self-report scale designed to detect the presence and severity of mild degrees of mood disorder, anxiety and depression in medical hospital patients, as well as in primary care patients. It consists of seven anxiety items (HADS-A) and seven depression items (HADS-D).

The CES-D scale (Hautzinger and Bailer, 1993) is a 20-item self-report scale which was developed to measure the level of depressive symptomatology in the general population and which placed particular emphasis on depressed mood.

The BDI (Hautzinger *et al.*, 1994) is a 21-item self-report inventory measuring the severity of depression. It addresses mood, cognitive and physical symptoms.

In all three questionnaires the time frame is 'during the past week'.

Patients' acceptance

In order to investigate patients' acceptance of the CAT administration, we asked respondents to rate 10 statements regarding the handling of the device and other cognitive and emotional aspects involved in computer-adaptive item administration. The statements were to be rated on a four-point scale, with two more negative and

two more positive response options. The original statements' content is reproduced in the results section.

Administration

All instruments, D-CAT and established questionnaires, were administered on personal digital assistants. Each item was presented separately on the personal digital assistant (PDA) screen. To select an answer, patients were instructed to use a pen. It was also possible to use the navigation button and some patients used this feature. Patients were shown how to use the PDA by a member of the nursing staff or a research assistant. The technical instruction was also summarized in text form on the first screen. Patients completed the questionnaires on their own in their room or in the community room of the ward. A member of the staff was constantly available in case any guidance was needed.

The time needed to complete the questionnaires – starting with the presentation of the first item – was recorded for all respondents.

Patients were informed about the aims of the study and gave their informed consent.

Data analysis

Whether or not patients from different diagnostic groups differed in their acceptance of the CAT was tested with Mann–Whitney-U-tests for non-parametrical data. Feasibility of the CAT was tested by measuring completion time and documenting patients' verbal reactions. Functioning was evaluated by recording item usage and their content and by evaluating measurement precision for each respondent. Convergent validity was determined by associations with established depression questionnaires (Pearson's correlations). To evaluate agreement between the D-CAT and established questionnaires, we carried out Bland–Altmann scatterplots (Bland and Altman, 1986). In this graphical method the differences between two assessment scales are plotted against the averages of the two assessment scales. A confidence interval is calculated in which approximately 95% of the differences should lie (mean of differences ±2 standard deviations). Feasibility, functioning and convergent validity were tested in sample A.

Discriminant validity was tested by comparing depression scores between diagnostic groups for the D-CAT, as well as with the established questionnaires. In sample A, patients with a depression were compared with patients with a mental or behavioral diagnosis other than depression and with patients with no ICD-10 F diagnosis using *t*-tests. In the sub-samples that had completed the D-CAT along with established questionnaires, namely the HADS and the CESD scale ($n = 127$), or respectively the BDI ($n = 111$), comparisons were made between different diagnostic groups, i.e. depression, adjustment disorders, anxiety disorders, other mental or behavioral disorders, or no diagnosis according to ICD-10 F. As some subsamples were relatively small, we applied Kruskal Wallis tests.

Discriminative power was additionally tested by univariate discriminant analyses (Efron, 1975). These were conducted in two samples: first in the clinically diagnosed sample (sample A); second in the CIDI-interview diagnosed sample (sample B). The analyses were conducted separately for all included instruments.

Results

Acceptance by the patients

The overall acceptance of the use of the personal digital assistant and of the mode of item presentation was fairly high. For nine out of ten questions on that issue more than 80% of the respondents chose a positive response option. The only major point of criticism was the size of the text on the screen, as 21% of the patients considered it to be too small (Figure 1).

None of the acceptance items was age-correlated or differed between age groups – not even when over 70-year-olds were tested against other age groups – nor were there any significant differences with respect to gender.

For three out of ten acceptance items we found differences between diagnostic groups. Patients diagnosed with a depression less often rated the handling of the device difficult than patients with other disorders ($U = 9869$, $p = 0.007$). They less often reported technical problems with the device ($U = 10{,}361$; $p = 0.010$) and they less often affirmed that using the computer disturbed them ($U = 10{,}423$; $p = 0.037$).

Measurement burden and precision

At a predetermined measurement precision of a SE of 0.32 or less (pertaining to a reliability of $r = 0.90$), the CAT produced an estimation of the latent trait across the whole range of the depression continuum, including the low and high extremes of the continuum. This level of reliability was reached after 4–18 items. On average, only six items were used (SD = 2.5 items). Figure 2 shows the mean number of items (and the SD) needed for a precisely estimated CAT score at different levels of the latent trait.

The reduction in respondent burden is in accordance with the speed of completing the D-CAT. Mean

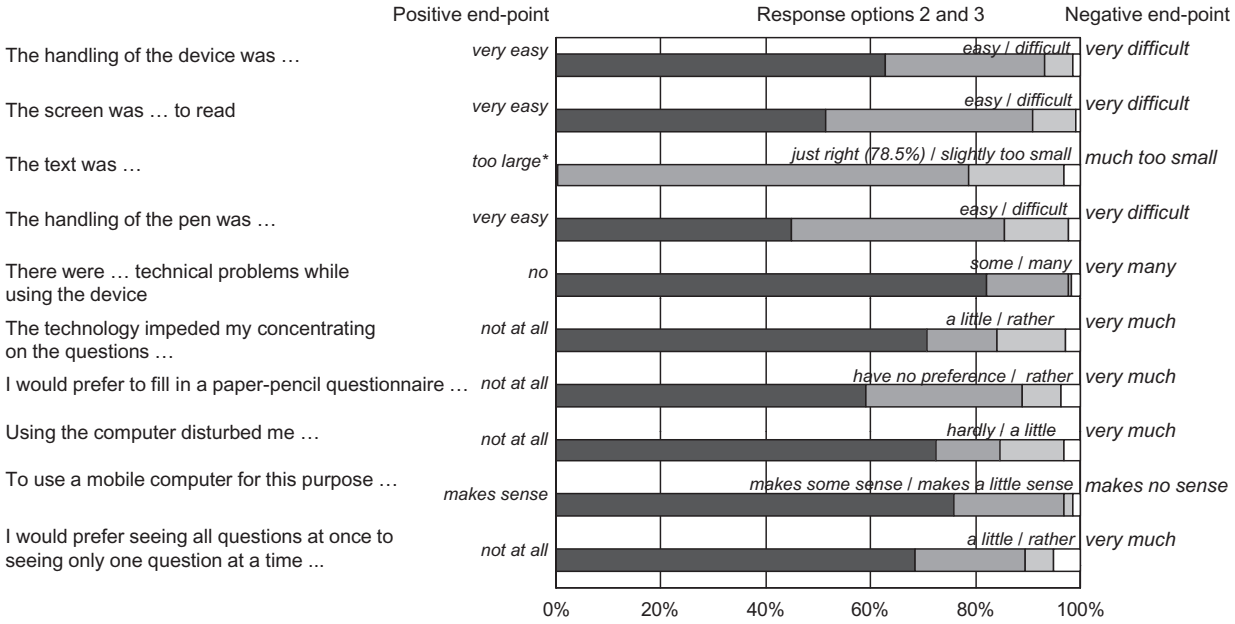

Figure 1 Patients acceptance of using the personal digital assistant (PDA) for computer-adaptive testing; response given included for each item (n = 423 respondents).

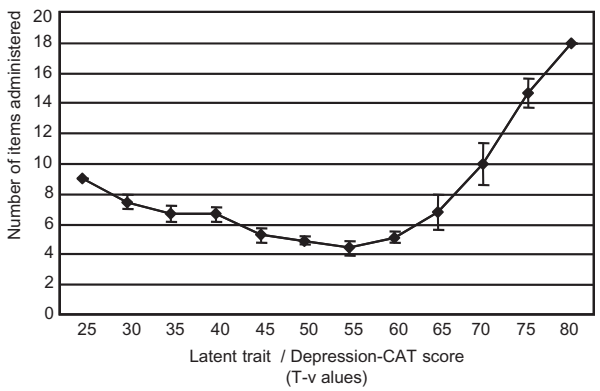

Figure 2 Number of items needed for a precise CAT score estimate (stopping rule SE < 0.32) as a function of the latent trait (T-values, M = 50, SD = 10); bars indicate the standard deviation (n = 423 respondents).

completion time was 1 minute 14 seconds (SD = 1:32). Completion time ranged from 4 seconds to 13 minutes. In 60 % of the cases it took respondents 1 minute or less to answer the CAT. In 95% of the cases it took them less than 4 minutes. As a point of comparison, the mean completion time for the HADS depression scale was 1 minute 57 seconds (SD = 1:58, range 1:1–13:6). The mean completion time for the CES-D scale was 3 minutes 4 seconds (SD = 2:30, range 0:39–13:24). The mean completion time for the BDI was 5 minutes 20 seconds (SD = 3:1, range 1:20–14:54).

Item usage

The real CAT application used 27 items (42%) from the pool (64 items), start item included (Table 1). Thirty-seven items were not used (Table 1). Apart from the first four items, no item was used in more than 10% of the cases. Four items accounted for 50% of all item administrations.

Concerning the 27 items that were used during the CAT application and the 37 items that remained unused, no mean differences in the number of response options or the location parameters on the latent trait continuum existed. However, unused items had a lower mean slope (1.02) than used items (1.33) (t = –3.66, df = 62, p = 0.001).

The fact that more than half of the items were not used has an impact on the represented content domains. This is not so much a question of statistically significant differences but of the content focus of the item set ultimately applied. In the set of used items, only the first content domain, i.e. depressed mood, is represented by more than two items. Three domains – concentration, suicidal thoughts and sleep disturbance – are covered by only one item. The last domain, loss of appetite, is not represented at all.

Table 1 Overall item usage of the D-CAT (left) and list of items not used by the D-CAT (right)

Used items (abbreviated content)	Subdomain	Unused items (abbreviated content)	Subdomain
Felt depressed (start item)	1	Things worried me	1
		Lack of concentration	6
Depressed	1	Everything is straining	3
Sorrowful	1	Talking less	2
Managing less/feeling down	2	Not able to pull myself together	2
Sad	1	Making oneself comfortable (R)	1
Not to cheer up	1	Being interested in something (R)	5
Feeling happy (R)	1	Self-acceptance (R)	4
Life is failure	1	Unsatisfied/bored	1
Desire to fall into sleep	8	Crying	1
Enjoying life (R)	5	Bad tempered	1
Unbearable sad/unhappy	1	Unable to work	2
Content (R)	1	Disturbed appetite	9
Inhibited/tense	1	Concentrated (R)	6
Future seems hopeless	1	Relaxed (R)	1
Impassive	1	Lethargic	2
Lack of interest	5	Cheerful (R)	1
Downhearted and sad	1	Feeling tired	3
Troubled	1	Worried	1
Thoughts of suicide	7	Insecure	1
Being easily fatigued	3	Quickly exhausted	3
Problems in decision making	6	Displeased with abilities	4
Feeling of being punished	4	Feeling on verge of breaking down	1
Despair/panic	1	Feeling unreal	1
Full of energy (R)	2	Feeling empty, paralyzed	3
Weary	3	Feeling ashamed when can't do something	4
Self-confident (R)	4	Socially impaired	2
Even-tempered (R)	1	Overwrought	3
		Happy (R)	1
		Feeling safe (R)	1
		Worried something will go wrong	1
		Feeling well (R)	1
		Glad (R)	1
		Jolly (R)	1
		Optimistic (R)	1
		Things won't go my way	4
		Rely on coping abilities (R)	4

Note: R = reversed scoring.
Subdomains: (1) Depressed mood, (2) Activity disturbance, (3) Fatigue, (4) Self-reproach/guilt, (5) Loss of interest, (6) Poor concentration/indecisiveness, (7) Suicide ideation, (8) Sleep disturbance, (9) Appetite loss.

Correlations with established instruments – convergent validity

The Pearson correlations between the D-CAT scores and the sum scores from the established questionnaires for depression were $r = 0.72$ for the HADS-D scale, $r = 0.77$ for the CES-D scale, and $r = 0.68$ for the BDI. Although the results confirm the relation between CAT scores and scores from established questionnaires that was found in earlier simulation studies (Fliege *et al.*, 2005), the association is less pronounced.

The Bland–Altman scatterplots, seen in Figure 3, show that 95.2% of the cases for the D-CAT and the HADS lie within the 95% limit of agreement; 94.4% of the cases for the D-CAT and the CESD scale and 97.3% of the cases for the BDI attain this agreement. This suggests high

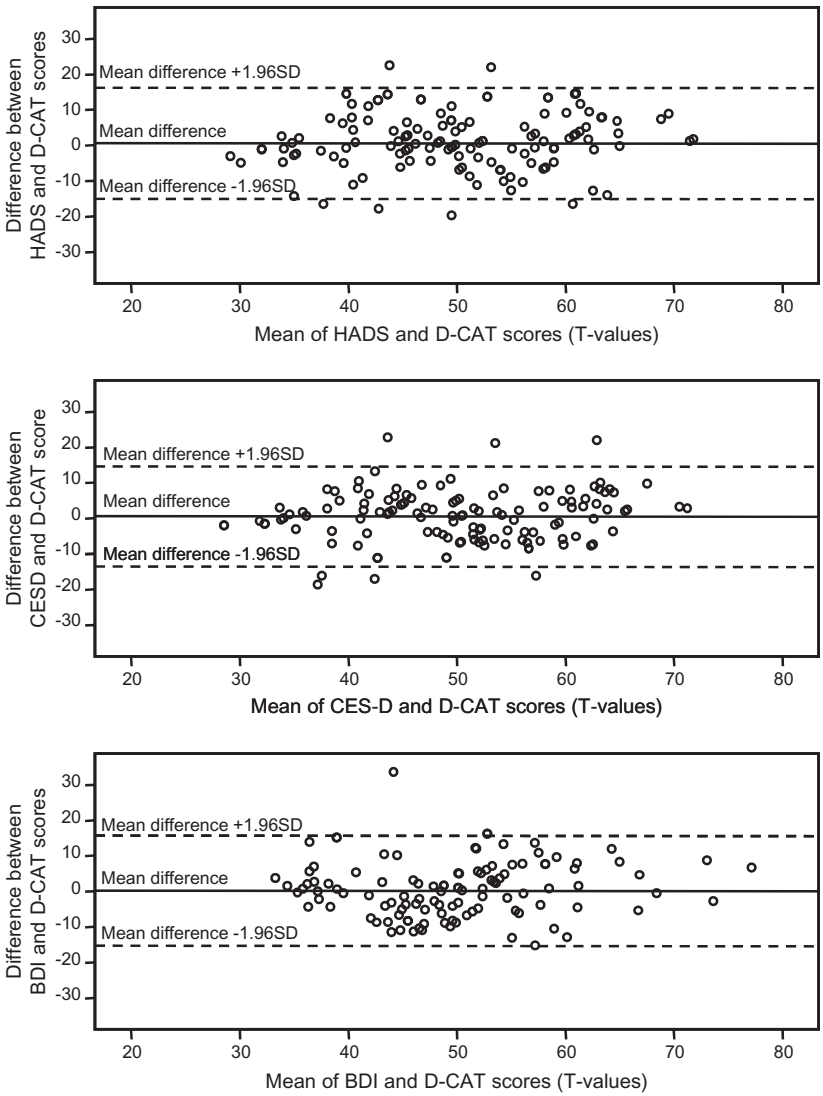

Figure 3 Bland-Altman scatter plots of the difference (y-axis) and the mean (x-axis) in depression scores between the Depression-CAT (6 ± 2.5 items; mean ± SD; stopping rule SE ≤ 0.32) and the HADS (top) and the CESD scale (middle) (both subsample A1, $n = 127$), and the BDI (bottom) (subsample A2, $n = 111$) with the 95% confidence interval; all values are T-transformed.

agreement between the D-CAT and the established depression instruments.

Discriminative validity

Patients clinically diagnosed with depression ($n = 78$) had a mean D-CAT score of 54.8. Patients with other mental or behavioral diagnoses ($n = 313$) had a mean score of 49.6 ($t = 3.9$, df = 107, $p < 0.001$) and patients with no mental or behavioral diagnosis ($n = 32$) had a mean score of 43.6 ($t = 5.51$, df = 113, $p < 0.001$).

The discriminative power of the D-CAT was tested along with that of the other instruments. Table 2 reports Wilk's lambda, an inverse measure for the distinction of the groups, the eigenvalue, a measure indicating the ratio of explained and unexplained variance, the chi-square statistics of group differences and the rate of correctly classified cases. Discriminant analyses yielded an equal power to discriminate between patients with and patients without a diagnosed depression for the D-CAT, the HADS, the CES-D scale and the BDI. This held true for clinically diagnosed patients and for those diagnosed by CIDI interviews.

Table 2 Univariate discriminant analyses of the D-CAT and the established depression inventories HADS, CES-D, and BDI, to discriminate between patients diagnosed with a depression and patients diagnosed with any ICD-10 F disorders other than depression

Instruments	Wilk's lambda	Eigenvalue	Chi-square	df	*p*	Correct classification
Clinical diagnoses (sample A1, n = *119)*						
D-CAT	0.91	0.10	10.7	1;115	0.001	62%
HADS-D	0.90	0.11	12.9	1;115	0.000	62%
CES-D	0.92	0.08	9.3	1;115	0.002	62%
CIDI diagnoses (sample B, n = *114)*						
D-CAT	0.88	0.13	14.0	1;112	0.000	60%
HADS-D	0.90	0.12	12.4	1;112	0.000	62%
BDI	0.89	0.12	12.8	1;112	0.000	59%

Note: Eight out of 127 patients in sample A1 had no ICD-10-F diagnosis and were therefore excluded from discriminant analyses. The CES-D scale was only administered in the clinical sample; the BDI was only administered in the interview sample.

In the sub-samples that included established depression questionnaires along with the D-CAT, Kruskal Wallis tests of differences between diagnostic groups were significant for all instruments (see Figure 4). Patients with a depressive disorder (n = 35, ICD-10 F32–34) had the highest D-CAT scores. Patients with adjustment disorders (n = 22, ICD-10 F43) had slightly lower scores, followed by patients with anxiety disorders (n = 12, ICD-10 F40–41) and patients with other mental or behavioral disorders (n = 50, ICD-10 F0 physiologically conditioned disorders, F10 substance abuse/addiction, F42 obsessive-compulsive disorders, F44 dissociative disorders, F45 somatoform disorders, F50 eating disorders, F6 personality disorders). Patients with no diagnosed mental or behavioral disorder (n = 8) exhibited the lowest D-CAT scores. According to the box-plots and the test statistics, the differences between diagnostic groups were similar for the D-CAT and for the HADS-D and the CES-D scale. Although the pattern of group differences was largely similar for the BDI, the chi square failed to be significant, probably due to the smaller size of this sub-sample.

Discussion

While some CATs for measuring mental health constructs such as depression, anxiety or perceived stress have recently been developed and tested in simulation studies (Fliege *et al.*, 2005; Gardner *et al.*, 2004; Walter *et al.*, 2007), reports of real applications of CATs in mental health contexts are still rare. Our study evaluated the real application of the pilot version of an IRT-based computerized-adaptive test for depression in a clinical setting. The items which the CAT applies are drawn from a subset of items from existing questionnaires used in the clinical diagnostic routine of the hospital. They were originally selected by clinical experts as indicative of the depression construct as defined by the DSM-IV. The items fulfill the statistical requirements of the IRT framework and have stood the test of simulation studies (Fliege *et al.*, 2005). In the real application study reported here, the D-CAT's feasibility, acceptance, economy and discriminative power was evaluated.

The application of the D-CAT was feasible and patients' overall acceptance of it was high. One of the few points of criticism was that 21% of the patients evaluated the size of the text on the screen as being too small. A future target of technical refinement – especially for a device in a clinical context – must be to optimize text size and readability on the PDA screen. Other occasional comments by patients referred to the handling of the pen. In fact, since the termination of the study phase, we have encouraged patients to use the navigation button instead. Patients may find this easier as most nowadays have experience with similar devices, such as remote controls or mobile phones. Remarkably, none of the acceptance items was age-dependant. So, problems with text size or handling of the pen, at least in our sample, could not be attributed to biological age or generational affiliation.

Nonetheless, generally speaking, the quality of electronic patient-reported outcome data depends to some extent on optimally manageable technical equipment. Issues like text size on the screen or difficulties in handling a pen must be identified at an early stage and communicated to developers of the technical devices.

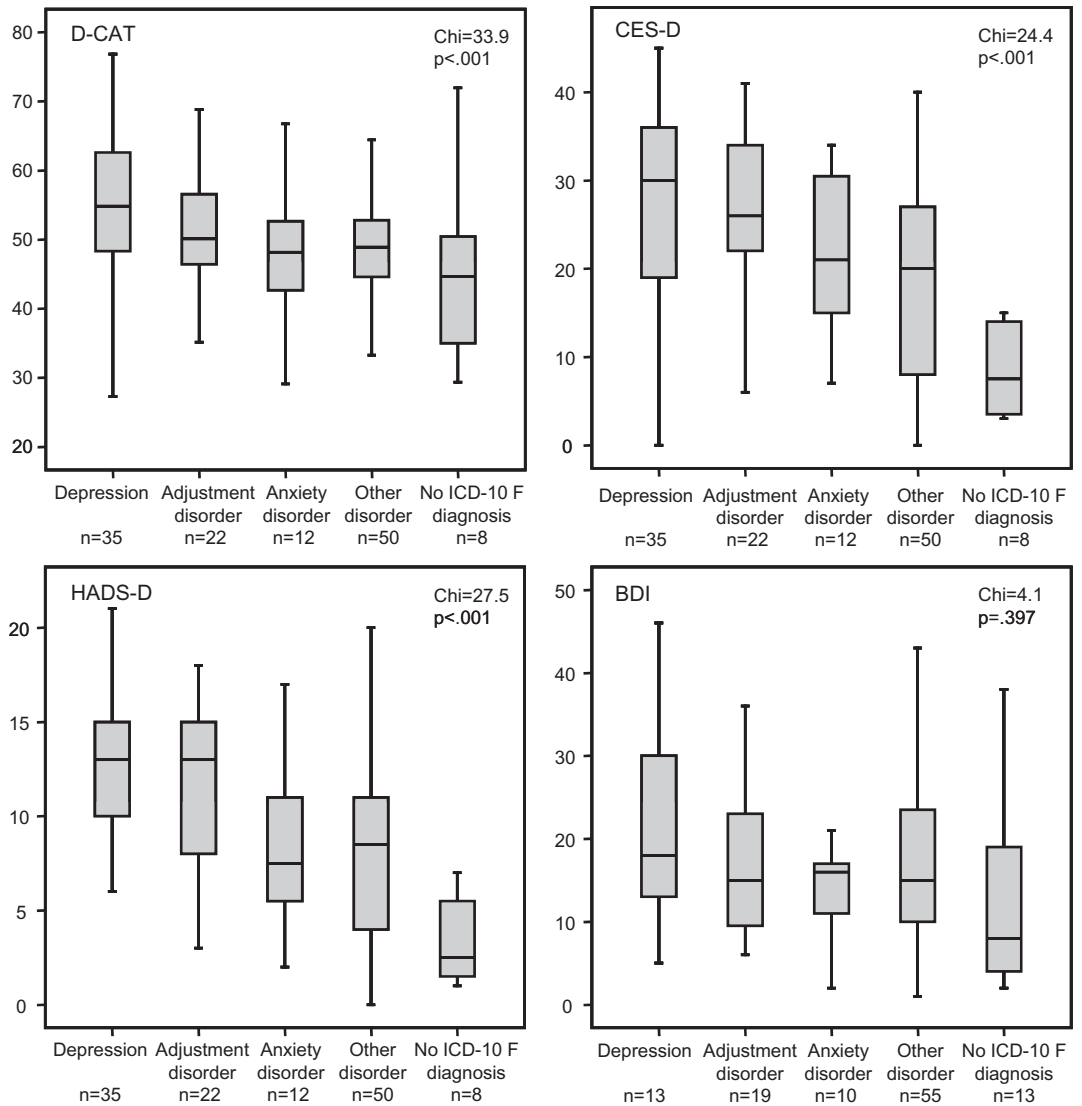

Figure 4 Depression scores for diagnostic groups by different measurement instruments (group differences: Kruskal Wallis test, df = 4).

The slightly more favorable acceptance ratings given by patients diagnosed with depression compared to patients with other mental or behavioral disorders may well speak of a lower readiness in depressed patients to express criticism. We would not interpret this as indicating a truly lower response burden in that group.

The D-CAT application performed an economic and, at the same time, reliable assessment across the whole range of the latent trait continuum. Though the time saved may seem small at first glance, it may well be significant in the long run, especially if the CAT approach is applied to other mental health constructs that are measured in a larger clinical routine context. On average, six items were needed for a precise score estimation, whereas a range of seven (HADS depression scale) to 21 (BDI) items is needed in the established questionnaires. Overall item savings reported in the literature on other CATs range from 50% to 92% (Bayliss *et al.*, 2003; Gardner *et al.*, 2004; Siebens *et al.*, 2004; Simms and Clark, 2005; Ware *et al.*, 2005).

Although content coverage was good, it is noteworthy that a large portion of the item bank was not used during the real CAT application. This is likely due to certain item characteristics, above all item information. Generally,

items with a higher slope parameter have a higher level of item information and items with a higher level of information are more likely to be chosen by the algorithm than items with a lower level of information.

In fact, in our CAT algorithm, items are selected by their level of information. Thus, it was not surprising that the items that had remained unused were those that, on average, had lower slope parameters. To select items by their level of information is a prerequisite for a reliable score estimation. A certain drawback of the item information criterion, however, is that item content will play no role. An inspection of the remaining items' content reveals that three domains are covered by only one item. In the case of the concentration domain, this is a consequence of item usage, whereas the domains sleep disturbance and suicide had initially been represented in the pool by only one item each. The last domain, appetite disturbance, is not represented at all. This does not impair a reliable estimation of the latent trait, as our results demonstrated. Theoretically, it would be possible to come to a reliable estimation of the latent trait for a subject after administration of only one or two items regardless of the content domains they cover. However, in our current version, the D-CAT begins with a fixed starter item dealing with depressed mood, a core facet of depression. Moreover, given unidimensionality of the construct as it had been secured in the first steps of CAT development, each item will be a valid indicator of the same latent trait. This CAT measures the continuum of depression, not its profile. Nevertheless, from a clinician's point of view, it may well be desirable to include certain item contents in any case.

In future works, item content will be investigated in greater depth. A possible strategy would be to use an algorithm that includes a content-balancing selection criterion. A content balanced item administration will also entail a minimum number of presented items. So far, we have not preset such a minimum. The minimum of four items needed for a precise estimation in our study was due to the specific set of items and their characteristics. Another strategy to elaborate and refine the present D-CAT in this respect is to generate new items and link them with the existing pool.

Another issue to be addressed in future research concerns the recall period and response format. Constructing an item bank from existing items as we did – and not starting from scratch – usually results in items with differing recall and response formats. The adaptive testing procedure means that recall period and response format could even switch with each new item presentation. Obviously, the large majority of participants in our study accepted this. However, it remains unclear if and how this influences answering behavior. While some authors even suggest that a switch of response formats may help to keep test takers' concentration high, a varying recall period is certainly not desirable. Therefore, one of the next steps of CAT development will be to standardize the recall period (best comparability with other instruments would be achieved by a period of one week). Another next step could be to adjust the present response formats to a common uniform response format for all items. Obviously this will entail testing the items' psychometric properties in a new sample.

Although the IRT framework has been successfully applied in order to refine established fixed-form questionnaires for depression, like the BDI (Bedi *et al.*, 2001; Kim *et al.*, 2002) or the CESD-scale (Chan *et al.*, 2004; Orlando *et al.*, 2000), there are only a few published reports on working CAT applications on any clinical construct, not to mention mental health constructs (Handel *et al.*, 1999; Walter *et al.*, 2007; Ware *et al.*, 2003). This may be due to the initially laborious and costly development. CAT development requires large initial samples (Edelen and Reeve, 2007), specific software development, and considerable effort in item analyses. However, there is good reason to assume that the application of CAT in the clinical routine will be more time-saving and economic in the long run than using static questionnaires. Yet, the point at which the investment in CAT development will actually pay off still needs to be demonstrated by cost-effectiveness studies

Though the routine clinical use of CATs is still in its infancy, a wider dissemination of CATs in health care will most likely occur within the next years due to a recently begun US nationwide initiative funded by the National Institute of Health (NIH) called Patient-Reported Outcomes Measurement Information System PROMIS (Fries *et al.*, 2005). PROMIS aims 'to revolutionize the way patient-reported outcome tools are selected and employed in clinical research and practice evaluation' by developing IRT-based CAT item banks for five central health domains (physical function, pain, fatigue, mental health, and role functioning). These CATs will be tested and validated across seven primary research sites led by a statistical coordinating centre and shall become publicly available in 2009.

With respect to the discriminative power of diagnostic groups, the mean D-CAT scores of patients with different mental or behavioral disorders showed a very similar pattern as those found in the established depression questionnaires. Discriminant analyses could demonstrate that the D-CAT has the same discriminative power as

established questionnaires. Yet, these findings are limited to patient samples. Future research on group discrimination of the D-CAT has to include mentally healthy samples in order to investigate diagnostic sensitivity and specificity. This will offer opportunities to develop cut-off scores and to relate CAT-scores to cutoffs from established questionnaires.

In summary, the CAT method was well accepted by the patients. The D-CAT achieves both an economic and precise measurement of depression, covering the whole continuum of the latent trait depression. The application under real conditions shows that depression is assessed validly in a similar way as established standardized questionnaires for depression do.

Acknowledgements

The study design was approved of by the Committee on Ethics of the Charité.

Parts of the results have been orally presented at the PROMIS Inaugural Conference in Gaithersburg, MD, September 11–13, 2006.

Parts of the study were funded by the German Research Foundation DFG (RO 2258/2–1). Principal investigator for this part was PD Dr Matthias Rose. The present phase was funded by the Charité Universitätsmedizin Berlin Research Fund (UFF-2006–088), with PD Dr Herbert Fliege as principal investigator for this phase. Current or former members of the study group are: Dr Janine Becker, PD Dr Herbert Fliege, Dipl.-Psych. Simone Getrost, Dipl.-Psych. Anne Grimm, Prof. Dr Burghard F. Klapp, Dr Rüya-Daniela Kocalevent, PD Dr Matthias Rose, Dr Otto Walter. Jacob Bjorner, MD, is scientific advisor and employed by QualityMetric Inc. PD Dr Matthias Rose is now employed by QualityMetric Inc and by the University Hospital Hamburg Eppendorf UKE. Dr Janine Becker was formerly employed by QualityMetric Inc. Dr Otto Walter is now employed by Münster University.

Declaration of interest statement

All authors certify that there are no conflicts of interest.

References

Allenby A., Matthews J., Beresford J., McLachlan S.A. (2002) The application of computer touch-screen technology in screening for psychosocial distress in an ambulatory oncology setting. *European Journal of Cancer Care*, **11**, 245–253, DOI: 10.1046/j.1365-2354.2002.00310.x

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV)*, American Psychiatric Association.

Attkisson C.C., Zich J.M. (1990) *Depression in Primary Care: Screening and Detection*. University Press.

Baer L., Jacobs D., Meszler-Reizes J. *et al.* (2000) Development of a brief screening instrument: The HANDS. *Psychotherapy and Psychosomatics*, **69**, 35–41, DOI: 10.1159/000012364

Bayliss M.S., Dewey J.E., Dunlap I. *et al.* (2003) A study of the feasibility of internet administration of a computerized health survey: the Headache Impact Test (HIT™). *Quality of Life Research*, **12**, 953–961, DOI: 10.1023/A:1026167214355

Bech P., Rasmussen N., Raabaek Olsen L., Noerholm V., Abildgaard W. (2001) The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *Journal of Affective Disorders*, **66**, 159–164, DOI: 10.1016/S0165-0327(00)00309-8

Beck A.T., Steer R.H. (2003) *Manual for the Beck Depression Inventory*, Psychological Corporation.

Bedi R.P., Maraun M.D., Chrisjohn R.D. (2001) A multi-sample item response theory analysis of the Beck Depression Inventory-1A. *Canadian Journal of Behavioural Science*, **33**, 176–187.

Bendtsen P., Timpka T. (1999) Acceptability of computerized self-report of alcohol habits: a patient perspective. *Alcohol and Alcoholism*, **34**, 575–580.

Bjorner J.B., Kosinski M., Ware J.E. (2003) Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the Headache Impact Test (HIT TM). *Quality of Life Research*, **12**, 913–933, DOI: 10.1023/A:1026163113446

Bland J.M., Altman D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **1**, 307–310.

Bock R.D., Mislevy R.J. (1982) Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, **12**, 261–280, DOI: 10.1177/014662168200600405

Carlson L.E., Speca M., Hagen N., Taenzer P. (2001) Computerized quality-of-life screening in a cancer pain clinic. *Journal of Palliative Care*, **17**, 46–52.

Chan K., Orlando M., Gosh-Dastidar B., Duan N., Sherbourne C. (2004) The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Medical Care*, **42**, 281–289, DOI: 10.1097/01.mlr.0000115632.78486.1f

Chou K.L. (2007) Reciprocal relationship between pain and depression in older adults. *Journal of Affective Disorders*, **102**, 115–123, DOI: 10.1016/j.jad.2006.12.013

de Denus S., Spinler S.A., Jessup M., Kao A. (2004) History of depression as a predictor of adverse outcomes in patients hospitalized for decompensated heart failure. *Pharmacotherapy*, **24**, 1306–1310.

Edelen M.O., Reeve B.B. (2007) Applying item response theory (IRT) modeling to questionnaire development,

evaluation, and refinement. *Quality of Life Research*, 16, Suppl 1, 5–18, DOI: 10.1007/s11136-007-9198-0

Efron B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**, 892–898.

Embretson S.E, Reise S.P. (2000) *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates.

Fliege H., Becker J., Walter O.B., Bjorner J.B., Klapp B.F., Rose M. (2005) Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, **14**, 2277–2291, DOI: 10.1007/s11136-005-6651-9

Fries J.F., Bruce B., Cella D. (2005) The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology*, **23**, S53–S57.

Gardner W., Shear K., Kelleher K.J. *et al.* (2004) Computerized adaptive measurement of depression: a simulation study. *BMC Psychiatry*, **6**, 4–13, DOI: 10.1186/1471-244X-4-13.

Gaynes B.N., Burns B.J., Tweed D.L., Erickson P. (2002) Depression and health-related quality of life. *Journal of Nervous and Mental Disease*, **190**, 799–806.

Gilbody S., House A., Sheldon T. (2001) Routinely administered questionnaires for depression and anxiety: a systematic review. *British Medical Journal*, **322**, 406–409, DOI: 10.1136/bmj.322.7283.406

Groenvold M., Petersen M.A., Idler E., Bjorner J.B., Fayers P.M., Mouridsen H.T. (2007) Psychological distress and fatigue predicted recurrence and survival in primary breast cancer patients. *Breast Cancer Research and Treatment*, **105**, 209–219, DOI: 10.1007/s10549-006-9447-x

Handel R.W., Ben Porath Y.S., Watt M. (1999) Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment*, **11**, 369–380.

Hautzinger M., Bailer M. (1993) *Allgemeine Depressionsskala. ADS. Testmappe mit Handanweisung*, Beltz.

Hautzinger M., Bailer M., Worall H., Keller F. (1994) *Beck-Depressions-Inventar. BDI. Testmappe mit Manual*, Huber.

Herrmann C., Buss U., Snaith R.P. (1995) *Hospital Anxiety and Depression Scale*, Deutsche Version, Huber.

Katon W., Ciechanowski P. (2002) Impact of major depression on chronic medical illness. *Journal of Psychosomatic Research*, **53**, 859–863, DOI: 10.1016/S0022-3999(02)00313-6

Katon W., Sullivan M.D. (1990) Depression and chronic medical illness. *Journal of Clinical Psychiatry*, **51**, 3–11.

Kim Y., Pilkonis P.A., Frank E., Thase M.E., Reynolds C.F. (2002) Differential functioning of the Beck Depression Inventory in late-life patients: use of item response theory. *Psychology and Aging*, **17**, 379–391, DOI: 10.1037//0882-7974.17.3.379

Kobak K.A., Greist J.H., Jefferson J.W., Katzelnick D.J. (1996) Computer-administered clinical rating scales. A review. *Psychopharmacology*, **127**, 291–301, DOI: 10.1007/s002130050089

Kroenke K., Spitzer R.L., Williams J.B. (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, **16**, 606–613, DOI: 10.1046/j.1525-1497.2001.016009606.x

Meijer R.R., Baneke J.J. (2004) Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods*, **9**, 354–368, DOI: 10.1037/1082-989X.9.3.354

Muraki E. (1992) A Generalized Partial Credit Model: application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159–176, DOI: 10.1177/014662169201600206

Olsen L., Jensen D., Noerholm V., Martiny K., Bech P. (2003) The internal and external validity of the Major Depression Inventory in measuring severity of the depressive states. *Psychological Medicine*, **33**, 351–356, DOI:10.1017/S0033291702006724

Orlando M., Sherbourne C., Thissen D. (2000) Summed-score linking using item response theory: application to depression measurement. *Psychological Assessment*, **12**, 354–359, DOI: 10.1037/1040-3590.12.3.354

Rose M., Walter O.B., Fliege H., Becker J., Hess V., Klapp B.F. (2002) 7 years of experience using personal digital assistants (PDA) for psychometric diagnostics in 6000 inpatients and polyclinic patients. In: *Mobile Computing in Medicine. Lecture Notes in Informatics – Proceedings* (eds Bludau H.B., Koop A.), GI-Edition.

Scott K.M., Bruffaerts R., Tsang A. *et al.* (2007) Depression-anxiety relationships with chronic physical conditions: results from the World Mental Health surveys. *Journal of Affective Disorders*, **103**, 113–120, DOI: 10.1016/j.jad.2007.01.015

Siebens H., Andres P.L., Pengsheng N., Coster W.J., Haley S.M. (2004) Measuring physical function in patients with complex medical and postsurgical conditions: a computer adaptive approach. *American Journal of Physical Medicine and Rehabilitation*, **84**, 741–748, DOI: 10.1097/01.phm.0000186274.08468.35

Simms L.J., Clark L.A. (2005) Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, **17**, 28–43, DOI: 10.1037/1040-3590.17.1.28

Stansbury J.P., Ried L.D., Velozo C.A. (2006) Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES-D) scale. *Journal of Personality Assessment*, **86**, 10–22, DOI: 10.1207/s15327752jpa8601_03

Walter O.B., Becker J., Bjorner J.B., Fliege H., Klapp B.F., Rose M. (2007) Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Quality of Life Research*, **16** Suppl 1, 143–155, DOI: 10.1007/s11136-007-9191-7

Walter O.B., Becker J., Fliege H. (2005) Entwicklungsschritte für einen computeradaptiven Test zur Erfassung von Angst (A-CAT) [Developmental steps for a computer-adaptive test for anxiety]. *Diagnostica*, **51**, 88–100, DOI: 10.1026/0012-1924.51.2.88

Ware J.E., Bjorner J.B., Kosinski M. (2000) Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Medical Care*, **38**, 1173–1182.

Ware J.E., Gandek B., Sinclair S.J., Bjorner J.B. (2005) Item response theory and computerized adaptive testing: implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology*, **50**, 71–78, DOI: 10.1037/0090-5550.50.1.71

Ware J.E., Kosinski M., Bjorner J.B. *et al.* (2003) Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research*, **12**, 935–952, DOI: 10.1023/A:1026115230284

Wittchen H.U., Pfister H. (1995) *DIA-X Expert System for Diagnosing Mental Disorders*, Swets.

World Health Organisation (WHO) (1997) *Composite International Diagnostic Interview* (CIDI), WHO.

Zigmond A.S., Snaith R.P. (1983) The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, **67**, 361–370.

Zung W. (1965) A self- rating depression scale. *Archives of General Psychiatry* **12**, 63–70.