Research Article

FUNCTIONING AND VALIDITY OF A COMPUTERIZED ADAPTIVE TEST TO MEASURE ANXIETY (A-CAT)

Janine Becker, Ph.D.,[1]* Herbert Fliege, Dr. rer nat,[1] Rüya-Daniela Kocalevent, Ph.D.,[1,2] Jakob B. Bjorner, M.D.,[3,4] Matthias Rose, M.D.,[3,4] Otto B. Walter, M.D.,[5] and Burghard F. Klapp, M.D.[1]

Background: *The aim of this study was to evaluate the Computerized Adaptive Test to measure anxiety (A-CAT), a patient-reported outcome questionnaire that uses computerized adaptive testing to measure anxiety.* **Methods:** *The A-CAT builds on an item bank of 50 items that has been built using conventional item analyses and item response theory analyses. The A-CAT was administered on Personal Digital Assistants to n = 357 patients diagnosed and treated at the department of Psychosomatic Medicine and Psychotherapy, Charité Berlin, Germany. For validation purposes, two subgroups of patients (n = 110 and 125) answered the A-CAT along with established anxiety and depression questionnaires.* **Results:** *The A-CAT was fast to complete (on average in 2 min, 38 s) and a precise item response theory based CAT score (reliability > .9) could be estimated after 4–41 items. On average, the CAT displayed 6 items (SD = 4.2). Convergent validity of the A-CAT was supported by correlations to existing tools (Hospital Anxiety and Depression Scale-A, Beck Anxiety Inventory, Berliner Stimmungs-Fragebogen A/D, and State Trait Anxiety Inventory: r = .56–.66); discriminant validity between diagnostic groups was higher for the A-CAT than for other anxiety measures.* **Conclusions:** *The German A-CAT is an efficient, reliable, and valid tool for assessing anxiety in patients suffering from anxiety disorders and other conditions with significant potential for initial assessment and long-term treatment monitoring. Future research directions are to explore content balancing of the item selection algorithm of the CAT, to norm the tool to a healthy sample, and to develop practical cutoff scores. Depression and Anxiety 25:E182–E194, 2008.* © 2008 Wiley-Liss, Inc.

Key words: *item response theory (IRT); computerized adaptive test (CAT); anxiety; measurement; questionnaire; validity*

INTRODUCTION

Anxiety is one of the most frequent mental disorders. Average life-time prevalence ranges between 17% worldwide[1–3] and 29% for the US[4,5] with substantial heterogeneity across studies. Four to 66% of patients

[1]Department of Psychosomatic Medicine and Psychotherapy, Charité Universitätsmedizin, Berlin, Germany
[2]AB Prävention und Gesundheitsforschung, FB Erziehungswissenschaften und Psychologie, Freie Universität, Berlin, Germany
[3]QualityMetric Incorporated (QM), Lincoln, Rhode Island
[4]Health Assessment Lab (HAL), Waltham, Massachusetts
[5]Institute of Psychology, University of Münster, Germany

Contract grant sponsor: Department of Psychosomatics and Psychotherapy, Charité Berlin, Humboldt University Hospital, Germany.

*Correspondence to: Janine Becker, Department of Psychosomatic Medicine and Psychotherapy, Charité Universitätsmedizin Berlin, Luisenstrasse 13 A, D–10117 Berlin, Germany. E-mail: janine.becker@charite.de

Received for publication 23 May 2007; Revised 19 October 2007; Accepted 8 January 2008

DOI 10.1002/da.20482
Published online 31 October 2008 in Wiley InterScience (www.interscience.wiley.com).

in primary-care settings have been reported to have at least one concurrent anxiety disorder in addition to a medical condition or depression.[6,7] Comorbidity of anxiety disorders in patients suffering from diabetes, cancer, cardiovascular disease, and irritable bowel syndrome ranges between 11 and 40%.[7–12] Anxiety disorders represent about 30% of total expenditures for mental illnesses, and health-care expenditure doubles when a comorbid mental illness like anxiety is present.[13] Further, studies have also shown that anxiety symptoms are predictive for the treatment outcome of other medical conditions.[14–20]

Thus, clinicians face a major challenge in recognizing, diagnosing, and treating anxiety syndromes.[6] A literature search using the key words "anxiety" and "test" or "questionnaire" or "inventory" in the titles of all articles stored in the databases Psyndex, PsycInfo, Psyndex, and PubMed between 1950 and 2006 identified that more than 50 questionnaires have been used over the past three decades to measure anxiety. The most popular anxiety questionnaires—defined by the number of articles found—are currently the State Trait Anxiety Inventory [STAI; 136 articles],[21] the Beck Anxiety Inventory [BAI; 40 articles],[22] the Hospital Anxiety and Depression Scale [HADS; 28 articles],[23] and the Zung Anxiety Scale [9 articles].[24] Almost all questionnaires have been developed on the basis of "classical test theory" [CTT][25] and are available as paper-and-pencil surveys. For such conventional questionnaires, a large number of items are usually needed particularly in test batteries applied in clinical settings to cover a wide range of constructs such as anxiety with a high measurement precision. This causes test developers to compromise between measurement precision and response burden when developing a tool. The combination of a modern measurement approach called item response theory [IRT][26–28] with computerized adaptive testing [CAT][29–31] technology has the potential to provide shorter questionnaires without compromising on measurement precision or test validity.

IRT techniques are based on a family of models,[26–28] which model a non-linear probabilistic relationship of an item response to the underlying latent trait. This approach differs from CTT, which assumes a deterministic relationship between items and a "true score" being measured together with an error term.

IRT employs a probabilistic function called item response category function, which is determined by two item parameters: the item difficulty (in IRT terms: location parameter) and the item discrimination (in IRT terms: slope parameter). The item response category functions can be plotted as item response curves. To help understand the IRT-modeling see the item response curves of an exemplary Computerized Adaptive Test to measure anxiety (A-CAT) item in Figure 1. The curves in the graph show the probability of responding using a specific response option in relation to the latent trait (*z*-scores). The more anxious a subject (high latent trait score), the more likely he/she responds to the question "Have you been anxious, worried or nervous during the past month?" with the response option "4" (very much), and the less likely he/she responds using the response option "1" (not at all).

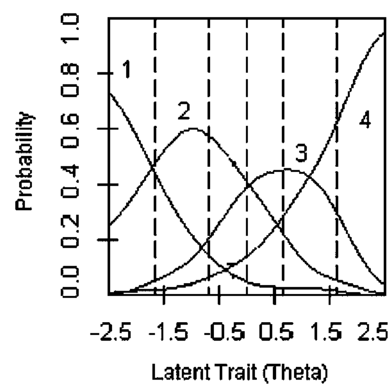

Figure 1. Item response curves of an exemplary computerized adaptive test to measure anxiety item.

IRT-modeling is useful for in-depth item analyses for test construction by evaluating the contribution of each item to overall test precision[32,99,100]and allowing comparison of item properties (like item precision and measurement range) across population subgroups (also called test of differential item functioning[33]). Further, IRT-modeling has been used to re-evaluate existing questionnaires like the Beck Depression Invetory[34,35] or the Hare Psychopathology Checklist.[36]

In CAT administrations IRT models enable selection of the most informative items for a particular range of anxiety and allow for estimation of comparable test scores from any combination of items along with individual assessment of measurement precision.[37] These features are not available by conventional (CTT) methods, because CTT assumes measurement precision to be the same across the measurement range. The assessment of measurement precision is critical for specifying CAT stopping rules. Finally, IRT methods allow for cross-calibrating different questionnaires measuring the same construct like anxiety on one common standard metric. This feature has been explored in several articles within the health field[32,38–43] and may have revolutionary effect in the field of measurement. Such metric—once established for scales measuring the same construct—would give clinicians a decision toolkit at hand for selecting those items/established questionnaires most appropriate for their sample [healthy, anxiety patients, etc].[42,44,45]

An IRT-based CAT selects only those items that are most informative for individual measurement combining the information given by previous item parameter estimations and the actual response of a patient to a question to choose this most informative/tailored item.[29–31] Thus, CATs provide the potential to realize a substantially shorter, less burdensome, and more precise measurement of anxiety.[46]

It should be mentioned that IRT is only one way of realizing a CAT, other ways are, for example, the countdown method described by Butcher.[47] This method classifies individuals into one of the two groups on the basis of whether they exceed or do not exceed a cutoff criterion on a given scale. If they reach a specific scale score after *a set of static items*, then a next set of static questions may be applied, if not the test ends. This method is less computational demanding, but also less adaptive than IRT-based CATs, the latter selecting items adaptively *after each item response*.

In 2004, our research group built a German A-CAT on the basis of IRT exploring some of the advantages noted above. The item bank of 50 items underlying the A-CAT had been built using items of a set of established questionnaires administered to 2,348 psychosomatic patients combining CTT and IRT-based methods for test construction as described elsewhere.[37,48] In previous studies the A-CAT had been tested for functioning and validity in small patient samples and simulation studies showing good reliability ($r>.9$), convergent ($r_{\text{A-CAT/HADS-A}} = .76$; $r_{\text{A-CAT/BAI}} = .55$; $r_{\text{A-CAT/NEO-PI-Neuroticism}} = .55$) and discriminative validity across diagnostic groups ($F = 35.6$, $df = 2$, $p \leq .001$).[49]

This study aims at evaluating the functioning and validity of the A-CAT in a large patient sample in clinical practice.

MATERIALS AND METHODS

SAMPLE

A sample of 357 consecutive patients treated at the department of psychosomatic medicine and psychotherapy, Charité Berlin, Germany, were administered the A-CAT. Data were collected between 07/2005 and 01/2006 during routine care. All patients were diagnosed by medical doctors specialized in internal medicine and psychotherapy with a background of several years of clinical experience. Patients constituting the sample were seeking health care in one of three settings: (a) in a psychosomatic medicine outpatient center (20.4%), (b) in a psychosomatic medicine liaison service at a general hospital (5.3%), or (c) during a psychosomatic medicine inpatient treatment (74.2%). They were approached by a nurse or an internee and asked, whether they were willing to participate in a study aiming at evaluating a new computerized test. Study consent had no implications on the treatment and no incentives were given.

Sociodemographic and diagnostic characteristics are summarized in Table 1. The sample is on average middle-aged (42 years), overrepresented by females (2/3), with most subjects being married (42%) or single, without a partner (25%). A third was employed (35%) while roughly another third was either retired (24%) or unemployed (14%).

Clinical diagnoses were given either after the outpatient or liaison visit or after a 2–8 weeks psychosomatic inpatient treatment (average: 19 days). Diagnoses were based on the clinical information gathered according to the criteria of the international classification of diseases [ICD-10 F][50] and supported by a diagnostic coding software (Diacos®, Berlin, Germany). Main clinical diagnoses according to ICD-10 F are illustrated in Table 1. About a fifth of the patients were diagnosed as having an adjustment disorder (21.6%) or a depressive disorder (18.8%). Other frequent diagnoses were somatoform disorders (16.0%), eating disorders (15.7%), or anxiety disorders (9.5%).

TABLE 1. Sociodemographics

Age (in years)	
Mean	42.6
SD	15.3
Range	18–76
Gender (in %)	
Female	68
Male	32
Family status (in %)	
Married	42.3
Single without partner	25.2
Single with partner	16.8
Divorced	12.9
Widowed	3.7
NA	0.6
Occupation (in %)	
Employee	34.8
Retired	24.4
Student/trainee	14.6
Unemployed	14.3
Self-employed	4.5
Housewife/man	3.4
Worker	2.2
NA	1.8
Clinical diagnoses ICD-10 F[a] (in %)	
F43 adjustment disorders	21.6
F3 depressive disorders	18.8
F45 somatoform disorders	16.0
F50 eating disorders	15.7
F40/41 anxiety disorders	9.5
F1 substance abuse/addiction	2.5
F44 dissociative [conversion] disorders	2.5
F6 disorders of personality and behavior	0.8
F42 obsessive–compulsive disorders	0.8
F0 disorders due to physiological conditions	0.3
F2 psychotic disorders	0.6
No F-diagnoses	10.9
Total (N)	357

SD: standard deviation. ICD: international classification of diseases. Clinical diagnoses were given after an outpatient or liaison, visit or after 2–8 weeks of psychosomatic inpatient treatment. Diagnoses listed are primary diagnoses of the patients.

There was a subgroup of patients having no primary F-diagnosis, but a somatic main diagnosis. It needs to be noted that there was an overlap in syndromes between the subgroups because most patients had one or more ancillary F-diagnoses.

A subsample of 125 inpatients completed the A-CAT and two established mood questionnaires in addition to the CAT: BAI,[22] and HADS.[23] Finally, out of the 125 patients, 110 patients also completed the Berlin Mood Questionnaire "Berliner Stimmungs-Fragebogen" [BSF][51] and the STAI.[21] The assignment to the subsamples was at random.

MEASURES

Anxiety-CAT (A-CAT). The A-CAT was administered drawing from an item bank of 50 items, which were the most informative for the individual taking the CAT. The item bank had been developed by re-analyzing 81 existing items given to $n = 2{,}348$ patients in a former study.[37] Re-analyses included the evaluation of item properties by confirmatory factor analysis, item response curves, and IRT-estimated item parameters].[52,53] Fifty items showing the best item properties were selected to build the A-CAT item bank. The final A-CAT item bank covers emotional (e.g. "being anxious"), cognitive (e.g. "being concerned"), and vegetative aspects (e.g. "being cramped") of anxiety.

First simulation studies of the A-CAT showed that anxiety could be estimated with 6.9 ± 2.6 items ($M \pm SD$), and the CAT algorithm having a higher discriminative power for patients at high and low levels of anxiety compared to conventional CTT-based sum scores [STAI].[37,48]

As illustrated in Figure 2 the A-CAT starts with the algorithm selecting and presenting the item (2) with the highest item information for the average score of the sample as the best initial score estimate (1). The first item given by the A-CAT is plotted in Figure 1. Then, the CAT algorithm uses the subject's response (3) to this item to estimate his/her CAT score including the CAT score precision (confidence interval) using the "expected a posteriori" method (4).[54] Once the CAT score is estimated, the CAT selects the next item based on the maximum information algorithm. This algorithm picks the item (2), which is most informative for the just estimated CAT score level using known item information parameters. After the next item administration (3), the CAT score and its precision are estimated again (4). The estimations are again used to pick the next most informative item and so forth (steps 2–4). The adaptive item selection and CAT score estimation is an iterative process stopping

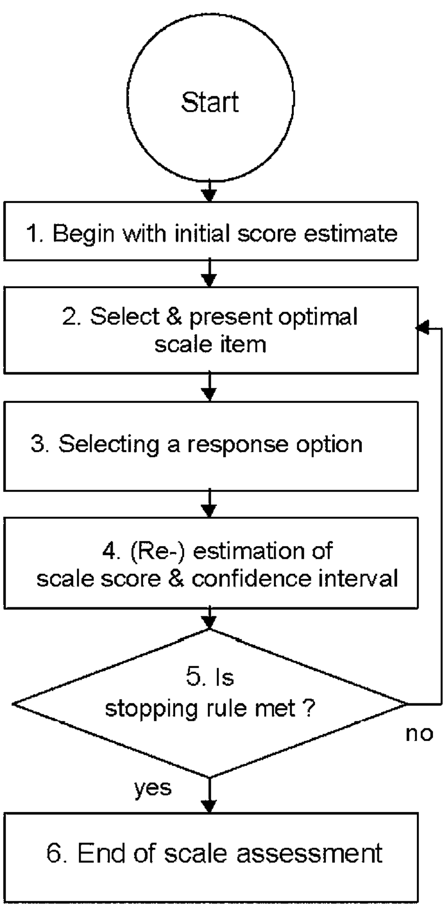

Figure 2. Flowchart of a computer adaptive test.

only when the individual CAT score precision reaches a pre-set target precision defined as the stopping criterion of the CAT (5). We decided to stop the test (6) when the standard error of measurement (*SE*) was at or below 0.32 *SD* units (equivalent to a reliability of >0.9). For further information on a CAT process, see Wainer.[31]

For illustrative purposes, see Figure 3 for screen shots of the A-CAT. The figure captures the first screen (instruction text of the A-CAT), two exemplary items ("anxious, concerned, or nervous"; "counterbalanced and self-confident") with chosen (highlighted) response options, and the last A-CAT screen ("thank you very much for completing the survey"). Please note that the A-CAT is in German and usually takes on average 6 items to complete.

Validation instruments. For validation instruments, the HADS, BAI, BSF, and STAI were given to subsamples. We chose those instruments to investigate how the A-CAT relates to conventional anxiety instruments differing in content and construct definition. In addition to the BSF, which is regularly administered for treatment monitoring at the hospital where the study was carried out, we chose the HADS and BAI due to their different content focus (BAI: somatic; HADS: anhedonic aspects of anxiety), and the STAI due to its wide use in research.

The HADS is a 14-item survey including an anxiety and a depression scale with 7 items. The anxiety scale covers somatic, cognitive, and emotional aspects of anxiety. Reliability for the anxiety scale has been estimated at .80 [Cronbach's α].[23] The BAI is a 21-item questionnaire, mostly assessing somatic symptoms of anxiety. Its reliability has been estimated from .85 to $>.90$.[55] The BSF is a 30-item survey comprising a 5-item scale of anxious/depressed mood in addition to five other scales (cheerful mood, engagement, anger, fatigue, apathy). The reliability of the subscale anxious/depressed mood is $r = .98$.[51] The STAI is a 40-item survey measuring anxiety as a more temporary state and/or more stable trait on two distinct, but empirically and conceptually overlapping scales. It has been mostly used for research purposes with a Cronbach's α of .88 (STAI-State) and .91 [STAI-Trait].[56]

Patients' acceptance. The acceptance of the A-CAT was tested by measuring the completion time of the survey for each patient and administering a 10-item patients' acceptance survey, which was developed by the authors to evaluate the technical handling of the device (5 items) and patients' opinion about using a computer device (5 items). Questions about the technical handling included items asking about difficulties reading the screen, handling the pen, or other technical issues, questions about the patients' opinion included items asking about the preference of a computer device over a paper–pencil survey, and whether the device had an impact on the concentration level. The items were displayed with four response options (1: very easy/not at all; 2: easy/a little; 3: difficult/some; 4: very difficult/very much).

DATA COLLECTION

All questionnaires were given on pocket PC's ("Personal Digital Assistants", PDAs), which have been implemented in the routine diagnostic procedure of our department since 1990.[57] The configuration of the PDA runs on Windows Mobile, the program language of the A-CAT is C++. They were either given to outpatients on their first visit while patients sit in the waiting area or to inpatients during their inpatient treatment. In the outpatient setting, the secretary sets the PDAs up with the patients ID, handed the PDAs to the patient and gives instructions for the survey completion, in the inpatient setting this job is performed by trainees/interns or nurses. On survey completion, the patients hand the pocket PC back and the secretary/trainee/intern or nurse plugs a cable into the PDAs connecting it to a stand-alone computer, which is used to transfer the survey data to the internal clinic network. The results of the questionnaire data were

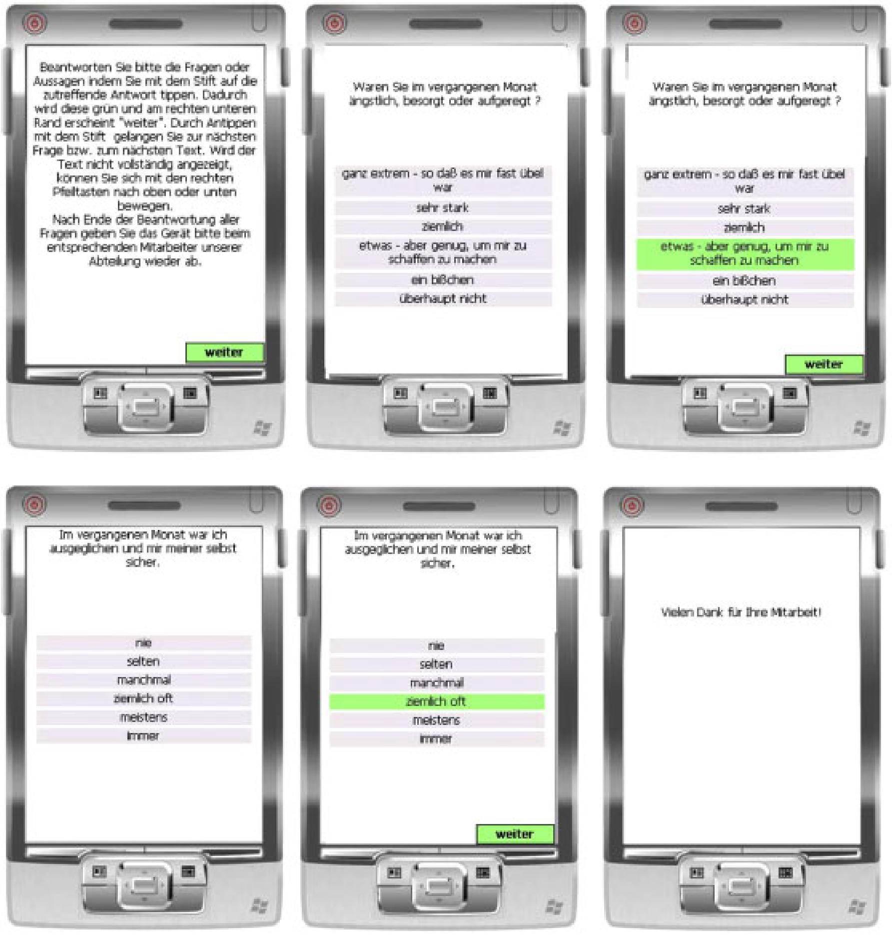

Figure 3. Screen shots of the computerized adaptive test to measure anxiety.

instantly reported on the screen, printed, and added to the patients' electronic medical record, which includes the physicians' diagnoses. This record is used by psychometrically trained clinicians to facilitate diagnosing by complementing the diagnostic interview at the start of any treatment and to support monitoring treatment outcome. For the purposes of this study, data were systematically retrieved from the clinic intranet and analyzed.

DATA ANALYSES

Functioning of the CAT was evaluated by records of the number and content of items displayed to the respondent as well as by inspecting the CAT score distribution. Measurement precision of each respondents CAT score was recorded (*SE*) and compared to the precision of the validation instruments (Cronbach's α). Acceptance of the A-CAT was evaluated by responses to the patients' acceptance survey. Response burden was explored by examining the completion time of the CAT compared to the other measures.

Convergent validation of the tool was performed by investigating the association between the A-CAT score and the sum scores of established instruments (HADS, BAI, BSF, and STAI) using scatter plots and correlational statistics (Pearsons' product moment correlation coefficient). Discriminant validity of the A-CAT was evaluated by analysis of variance statistics to test the mean score difference of the A-CAT score between patients with physician-diagnosed anxiety, mood or adjustment disorder, and patients with other mental (ICD-10 F-diagnosis) or medical disorders (no ICD-10 F-diagnosis). The latter one was included for comprehensiveness despite low clinical group sizes. Because comorbidity between anxiety and depression or adjustment disorders was high, a group called "anxiety and comorbidity" was included in the analyses. In addition, to specifically inspect pure differences between anxiety, depression, and adjustment disorders, those diagnostic groups were built without in-between comorbidity ("anxiety only," "depression only," and "adjustment disorder only"). The remaining categories of patients with further comorbidities (e.g. somatic disorders with mental disorders, etc.) were not included in the analyses to avoid confusion. Thus, diagnostic group sizes do not sum up to total sample sizes.

RESULTS

CAT FUNCTIONING: ITEM NUMBER, SCORE DISTRIBUTION, AND CONTENT

The A-CAT needed between 4 and 41 items to achieve the specified measurement precision (Fig. 4). On average 6 items were displayed ($SD = 4.2$ items).

A-CAT scores were transformed from IRT-based *z*-scores to *t*-scores ranging on a 0–100 scale with an average of 50 and a $SD = 10$. The A-CAT is scored in the direction that high scores mean high levels of

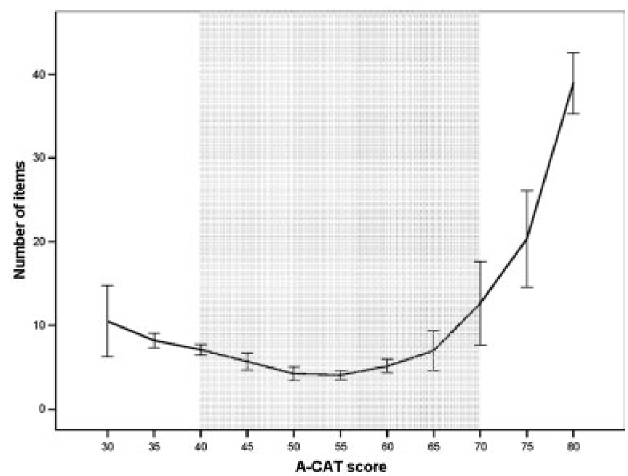

Figure 4. Number of items administered by the computerized adaptive test to measure anxiety as a function of the CAT score.

anxiety. About 90% of all patients completed 4–18 items (CAT score >40 and <70, see light gray shaded area in Fig. 4). Nine percent were given 4–9 items scoring in the lower (CAT score ≤ 40), and 2% answered 17–41 items scoring in the higher (i.e. more anxious) range (CAT score ≥ 70).

The A-CAT score average for the psychosomatic patients investigated here is 52.5 ($SD = 8.4$). The score distribution was skewed to the right with 75% scoring higher than 49 (75 percentile), i.e. most psychosomatic patients are more anxious than the average. Cut-scores for clinical meaningful interpretation beyond the norm-based comparisons need to be developed yet.

Four items accounted for 45% of all item administrations. Those items displayed in Table 2 asked for emotional aspects of anxiety such as "being anxious, worried, or nervous", "being counterbalanced", "being driven by anxiety and trouble," and "feeling secure". Among all items presented, half of them were reversed scored (r.) asking for "being counterbalanced and self-confident", "feeling secure," or "calm".

MEASUREMENT PRECISION

In our study the standard error of all CAT scores was on average $SE = 0.30$ ranging steadily on a low level between $SE = .27$ and .32. This translates into a reliability range between .93 and .95. For comparison of reliability, Cronbach's α of the other validation instruments calculated using the study data here was lower ranging between .83 and .93 (HADS-A: .83; BAI: .93; BSF-AD: .87; STAI-S: .90; STAI-T: .89).

ACCEPTANCE

The CAT process was well accepted by patients. For 9 out of 10 questions on the patients' acceptance of the device, 80% of the subjects chose a positive response option: they perceived the handling of the mobile computer/pen, and the readability of the screen as "very easy" to "easy". More than 80% had no technical problems using the device. The only major point of criticism on the device was that 21% of the patients thought the font size of the text on the screen was "too small". About 60% responded that they would prefer the computer survey over a paper–pencil survey, more than 30% said that they have no preference, and 10% would have preferred a paper–pencil survey. Eighty-five percent responded that the computer "did not" or "hardly" disturbed their concentration, whereas 12% said that they were "a little" disturbed, and 3% said they were "very" disturbed.[101]

RESPONSE BURDEN

Overall, the CAT survey was fast to complete. Respondents took on average 2 min, 38 s to complete the survey. For comparison the completion of the validation instruments were STAI (40 items): 4 min, 49 s; BAI (21 items): 2 min 47 s; HADS (14 items): 3 min 24 s; BSF (30 items): 3 min 26 s (for HADS and BSF only completion times of the whole scale including the anxiety subscale were recorded).

VALIDITY

Convergent validity of the A-CAT was supported by moderate correlations to existing tools as illustrated in Figure 5. The A-CAT correlated the highest with the STAI-S ($r = .66^{**}$). Close inspections of the scatter plots between established anxiety measures and the A-CAT revealed a substantial amount of variance of scores. Correlations between the A-CAT to discriminant constructs as measured by the remaining five scales of the BSF (cheerful mood, engagement, anger, fatigue, apathy) range between: $r_{\text{(BSF-cheerful mood/A-CAT)}} = -.47$ and $r_{\text{(BSF-anger/A-CAT)}} = .39$. Those results are in line with discriminant correlations of the other anxiety tools and the BSF scales ranging between $r_{\text{(BSF-cheerful mood/HADS-A)}} = -.63$ and $r_{\text{(BSF-anger/STAI-S)}} = .49$.

Discriminant validity of the A-CAT was overall somewhat better than for the other instruments (see Fig. 6). Not surprisingly, patients in the "anxiety and comorbidity" group had the highest anxiety scores in all validation instruments investigated: $M_{\text{A-CAT}} = 58.2$ ($\pm SD = 6.9$); $M_{\text{HADS-A}} = 58.3$ ($\pm SD = 19.4$); $M_{\text{BAI}} = 57.5$ ($\pm SD = 19.9$); $M_{\text{BSF-AD}} = 53.0$ ($\pm SD = 19.9$); $M_{\text{STAI-S}} = 65.5$ ($\pm SD = 14.9$); $M_{\text{STAI-T}} = 53.8$ ($\pm SD = 14.1$). The A-CAT seemed good in discriminating between patients with an anxiety diagnosis ($M = 58.2 \pm SD = 6.9$) and patients without a F-diagnosis, i.e. with somatic diagnoses only ($M = 41.9 \pm SD = 10.7$, $P < .001$). The mean differences for the anxiety disorders and somatic diagnoses only group were significant for the A-CAT ($P < .001$), the STAI scales (STAI-S: $P = .009$, STAI-T: $P = .011$), and the BSF-AD ($P = 0.002$), but non-significant for the other measures (HADS-A: $P = .902$; BAI: $P = .463$).

No anxiety instrument showed huge differences between clinical groups of anxiety only and depression

TABLE 2. Overall item usage of the A-CAT (left) and list of items not used by the A-CAT (right)

Used items			Unused items	
Abbreviated content	Subdomain	Percent	Abbreviated content	Subdomain
Being anxious, worried or nervous	E	15.8	Having fear	E
Being counterbalanced and self-confident (r.)	E	10.8	Feeling insecure	E
Driven by anxiety and trouble	E	10.2	Being afraid, sth. will go wrong	E
Feeling secure (r.)	E	7.9	Feeling insecure in groups	E
Being calm and even-tempered (r.)	E	5.8	Keeping calm in the face of problems (r.)	E
Looking on the black side causes panic	C	5.5	Being calm (r.)	E
Being excited	E	4.6	Feeling counterbalanced (r.)	E
Being strained	V	4.3	Feeling self-confident (r.)	E
Being relaxed or agitated (r.)	V	4.0	Feeling calm (r.)	E
Feeling nervous	V	3.6	Crowds scare me	E
Being light-hearted (r.)	E	3.0	Feeling secure and protected (r.)	E
Feeling relaxed (r.)	E	2.9	Feeling of not existing	E
Feeling released (r.)	E	2.1	Feeling like a stranger	E
Feeling concerned	C	1.5	Feeling well (r.)	E
Complaints due to inner fear	V	1.4	Being frightened of the future	E
			Being worried	C
			Being concerned	C
			Feeling worried	C
			Being afraid of not achieving goals	C
			Having lots of trouble	C
			Being concerned about one's health	C
			Feeling antsy	V
			Being fidgety	V
			Feeling tense	V
			Being cramped	V
			Being overwrought	V
			Feeling numb	V
			Able to making oneself comfortable/relax (r.)	V
			Being released (r.)	V
			Being nervous	V
			Feeling tense	V
			Body seems strange	V
			Problems to relax	V
			Lump in throat, pokiness or choking	V
			Being nervous	V

Subdomains: E: emotional, C: cognitive, and V: vegetative aspects of anxiety; (r.): reversed scoring.

only ($M_{\text{A-CAT}} = 52.3/53.5$, $M_{\text{HADS-A}} = 46.2/46.1$, $M_{\text{BAI}} = 38.6/39.1$). The mean differences for the anxiety versus depression groups were not significant for all measures (A-CAT: $P = .468$; STAI-S: $P = .193$; STAI-T: $P = .404$; HADS-A: $P = .991$; BAI: $P = .950$).

For some instruments (BAI, BSF, STAI), the "depression only" group scored slightly higher on the anxiety scales than the "anxiety only" group ($M_{\text{BSF}} = 54.4/39.2$, $M_{\text{STAI-S}} = 62.7/50.3$, $M_{\text{STAI-T}} = 57.6/49.7$). However, as expected the anxiety group scored higher (i.e. more anxious) on the A-CAT than the depression group ($M_{\text{A-CAT}} = 58.2/53.5$); thus supporting the discriminant validity of the A-CAT.

DISCUSSION

This study investigated the functioning and validity of one of the first IRT-based mental health CATs for clinical practice: the Anxiety-CAT.[37] Major findings of this study were that (A) the A-CAT was functioning well and accepted among patients, (B) response burden was low, and (C) validity was comparable to or better than established anxiety questionnaires.

(A) The A-CAT functioned well and was favorably perceived by the patients. As expected, the CAT algorithm selected the most informative items for each level and calculated IRT-based test scores and test takers precision. This led to a reduction in the number of items displayed (on average 6 items), while maintaining high measurement precision (reliability >.9). Eighty percent of respondents perceived the A-CAT in 9 out of 10 questions very positive, responding that the handling of the mobile computer/pen, and the readability of the screen were "very easy" to "easy." The CAT acceptance is in line with previous literature on computerized questionnaires reporting increasing

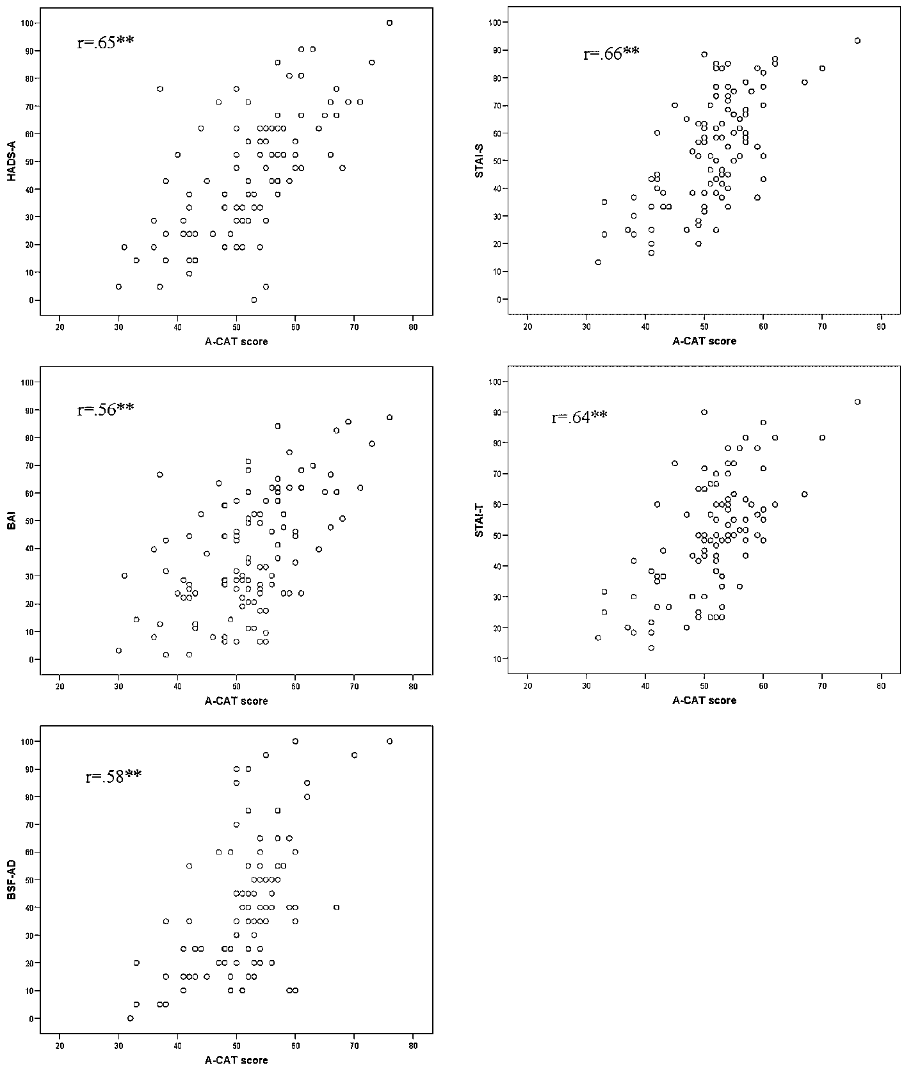

Figure 5. Scatter plots of the relation of the CAT score to established anxiety measures.

popularity of computerized tools.[57–59] Several authors demonstrate that they are less time consuming, more efficient, and very feasible of fitting into a routine clinical work flow.[60–62] Good acceptance of the A-CAT is also in line with a number of studies on the acceptance and even preference over paper–and-pencil surveys.[63–66]

(B) Response burden of the A-CAT was low. Although most anxiety questionnaires consist of 7–40 items, the A-CAT administered on average only 6 items similar to anxiety subscales of the HADS (7 items) and the BSF (5 items). This is in line with previous CAT studies reporting average test length reduction to 5[46] to 10 items[67,68] without a substantial loss of information compared to full-length CAT-item banks holding 23–71 items.

Concerning completion time, the A-CAT on average saved half the time that is needed to fill out a more extensive questionnaire like the STAI. However, no substantial time-saving compared to the shorter scales such as the BAI or the HADS is to be expected.

Overall the CAT literature reports item savings ranging between 50 and 92%,[63,69,70] and time-savings compared to full-length questionnaires ranging between 21 and 83% [Diabetes CAT, Osteoarthritis CAT, Headache CAT, PEDI-CAT: Pediatric Evaluation of Disability Inventory CAT, CKD-CAT: Chronic Kidney Disease.[68,71,72] High response burden in CATs mainly occurs at the extremes of the range, when there is a lack of informative items. The A-CAT compensates for this by administering more items. If more items were developed that were particularly relevant for very high or very low anxiety, the total response burden for the A-CAT could be diminished. Alternatively, the stopping rules could be made more flexible by criteria combining precision and number of items.

(C) Results suggest satisfactory content, convergent, and discriminant validity of the A-CAT. The item bank

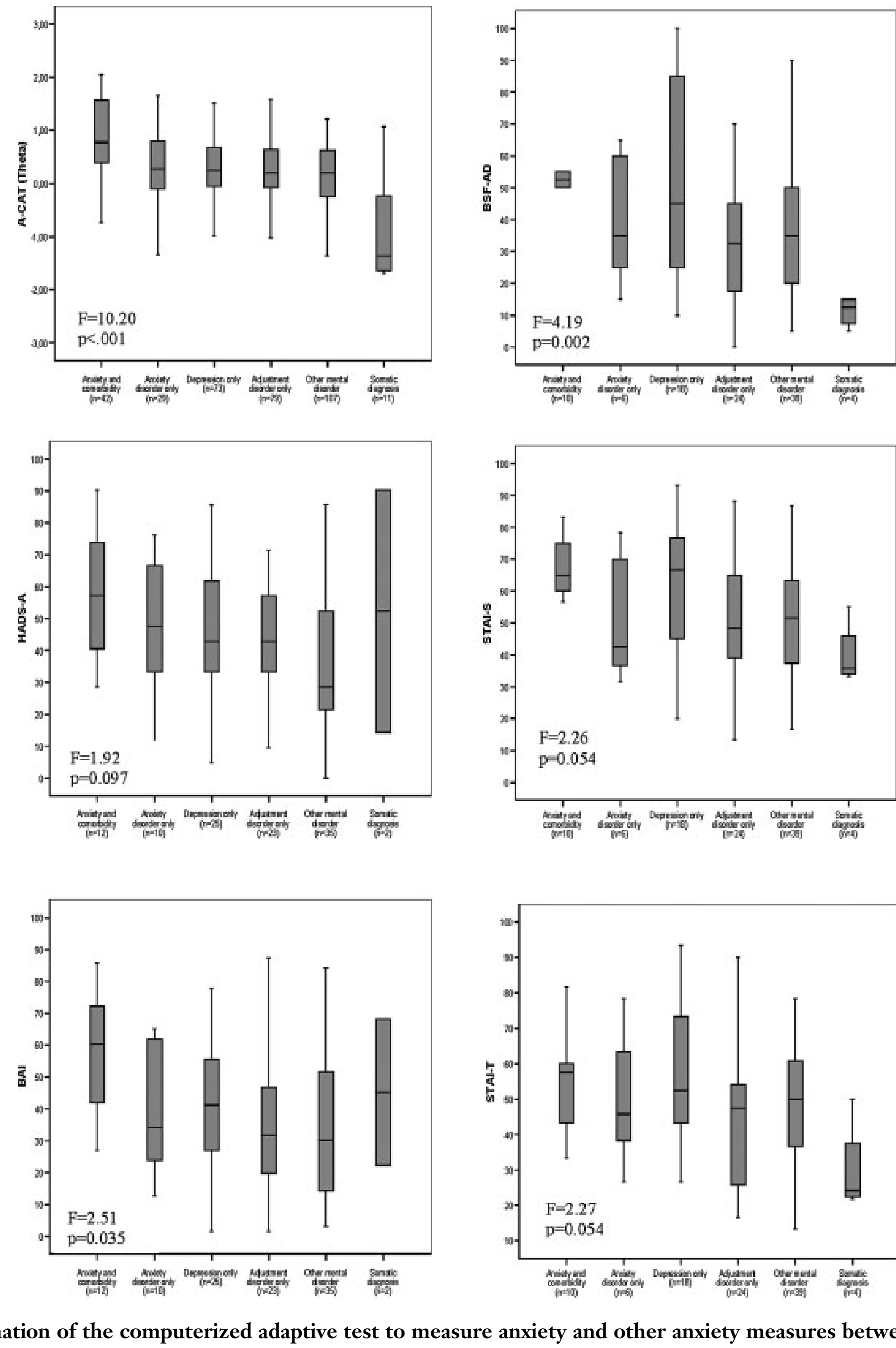

Figure 6. Discrimination of the computerized adaptive test to measure anxiety and other anxiety measures between diagnostic groups.

covers emotional, cognitive, and vegetative aspects of anxiety. Most frequently the A-CAT presented items asking for emotional aspects of anxiety-one of four aspects formulated in the four-factor model proposed by Beck et al.[22] It may need to be discussed, whether items asking for cognitive or vegetative aspects of anxiety should be displayed more often to counter-balance the frequent display of items assessing mainly the emotional aspect of anxiety to increase content validity. This can be achieved by using content-balancing item selection rules,[31,73] and by adding new items to the item pool, two directions we will take up in our work.

Convergent validity was indicated by moderate correlations to existing anxiety measures. The correlation of the A-CAT to the STAI was the highest, most likely because the A-CAT includes items similar to the STAI; whereas the correlation to the BAI was the lowest, most likely because 13 out of 21 BAI items assess physiological/somatic symptoms of anxiety,[74] which are not covered by the A-CAT. The BAI has been criticized for over-emphasizing panic attack symptoms.[75,76] From our study results, we may conclude that the A-CAT (like the STAI) may be more responsive to cognitive and affective components of anxiety, whereas the BAI may be more responsive to somatic components of anxiety. That may imply that patients, who are able to communicate emotional and cognitive aspects of anxiety may benefit more from the A-CAT, whereas patients, who have difficulties reflecting and communicating emotions or may tend to somatize them (for example, patients with somatoform disorders) may benefit more from the BAI. However, this needs to be further explored. We are currently discussing whether to cover somatic symptoms of anxiety by a separate CAT. When building the A-CAT item bank,[37] items asking for specific somatic symptoms of anxiety like sweating, flushing, trembling, dyspnea, problems swallowing, pain in breast or stomach, diarrhea or obstipation, tachycardia, dizziness, sleep disturbances, weakness or hot flashes needed to be excluded to fit unidimensionality.[1][48] It is an ongoing discussion about what degree of unidimensionality for fitting IRT-models is sufficient, and a few authors[77] discuss whether unidimensional IRT-modeling may not fare well with mental constructs being more frequently conceptualized as multidimensional constructs. Exploring multidimensional IRT-models is a promising way for further research in this field.[58]

Discriminant validity of the A-CAT was good in terms of differentiating between patients with anxiety and those without a mental diagnosis (no ICD-10 F). Discrimination between groups differing in mental disorders was not great, but still better than for other measures. This relates to a wider discussion about the justification of a general distinction between those concepts, which is questioned by high comorbidities between anxiety and depression disorders.[78–80] Alternative conceptualizations are discussed within the tripartite model, which assumes one global negative affective factor overlapping anxiety and depression, plus two specific factors, namely one specific to panic attacks, i.e. more vegetative anxiety symptoms, and one specific to depression, i.e. lack of positive mood (anhedonia) and hopelessness.[81–83]

Further, it was not surprising that the depressed patients scored higher on many of the anxiety measures, as several of these measures tend to assess general demoralization, rather than anxiety, specifically. That the A-CAT scores were still slightly higher for the anxiety group compared to the depression group may be an indication that adaptive procedures yield a more accurate picture than full-scale administration.

The principal limitation of our study is that so far only psychosomatic patients were included, thus data collection of healthy subjects or those who visit GP offices is needed. Another limitation is the small sample sizes for some diagnostic groups, calling for further replication of results. Though the routine clinical use of CATs is still rare, a wide dissemination will most likely occur within the next years due to a US nationwide initiative funded by the National Institute of Health called Patient-Reported Outcomes Measurement Information System.[45,84] Patient-Reported Outcomes Measurement Information System aims "to revolutionize the way patient-reported outcome tools are selected and employed in clinical research and practice evaluation" (www.nihpromis.org) by developing IRT-based CAT item banks for five central health domains: mental health, physical functioning, pain, fatigue, and role functioning. Those CATs will be tested and validated across seven US primary research sites led by a statistical coordinating center and become publicly available in 2009.

Following the successful A-CAT development, our group has more recently also built CATs for measuring depression[85] and stress perception and reaction.[86] Although IRT-based CATs have been implemented in large-scale ability testings for decades[67,87] [for example, SAT: www.collegeboard.com], applying CAT to clinical measurement is a fairly new scientific effort. Since 2000 other research groups shared this effort to develop CATs measuring (a) mental health,[58] (b) personality traits [Minnesota Multiphasic Personality Inventory-2:[77,88,89] NEO Personality Inventory-Revised:[42] Schedule for Nonadaptive and Adaptive Personality,[63] (c) quality of life impact of headaches, osteoarthritis, and fatigue among cancer patients [Headache Impact Test:[53,69] Osteoarthritis Impact CAT][71,90], and (d) physical functioning[91,92] [Mob-CAT][93] among others.

One of the most recent CAT developments in the field of personality testing is the CAT built Forbey and Ben-Porath.[77] In contrast to our study, they used the countdown method to explore two computerized adaptive versions of the Minnesota Multiphasic Personality Inventory-2. Similar to our study, they report substantial item and time-savings as well as external criterion validity of both CAT versions. In addition, they showed score comparability to the full-length scales (anxiety/psychasthenia scale: $r = .66$–$r = .82$), which we have shown for the A-CAT in previous simulation studies [$r = .97$].[37] Their study is among the first larger studies ($n = 517$) supporting reliability and validity of a CAT for personality testing.

The absence of similar large-scale clinical studies on CATs and new theoretical questions on unidimensionality and item fit being posed by IRT and CAT technology contribute to several authors questioning the appropriateness of adopting IRT and CATs for measuring mental health or personality testing. More large-scale validation studies on IRT-based CATs in

[1]Fit indices for a one-factor model: Comparitive Fit Index = 0.77–0.78, Tucker–Lewis Index = 0.75–0.76, Root-Mean-Square Approximation = 0.10; this fit is only moderate applying cutoff criteria of fit indices.[97,98]

mental health/personality measurement are needed to advance this field.

Overall, our study suggests that the A-CAT is a short, precise, and valid tool for assessing anxiety in patients suffering from anxiety disorders and/or other medical conditions. It holds the potential for routine screening and monitoring to improve the recognition of anxiety disorder in clinical settings,[94] for improving doctor–-patient communication,[95,96] tailoring treatment, facilitating referral to specialists, and monitoring outcome.

Future research directions include exploring techniques for content-balancing the item selection algorithm, developing healthy norms, and practical cutoff scores of the A-CAT.

Practical challenges remaining are the integration of CATs into comprehensive IT systems in hospitals and training clinicians to apply and interpret CAT scores in daily clinical routine.

Acknowledgments. We especially thank all patients and colleagues at the Department of Psychosomatics and Psychotherapy, Charité Berlin, Humboldt University Hospital, Germany, who helped in realizing this project.

REFERENCES

1. Ohayon MM. Anxiety disorders: prevalence, comorbidity and outcomes. J Psychiatr Res 2006;40:475–476.
2. Somers JM, Goldner EM, Waraich P, Hsu L. Prevalence and incidence studies of anxiety disorders: a systematic review of the literature. Can J Psychiatry 2006;51:100–113.
3. Starcevic V. Review: worldwide lifetime prevalence of anxiety disorders is 16.6%, with considerable heterogeneity between studies. Evid Based Ment Health 2006;9:115.
4. Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62:617–627.
5. Narrow WE. The numbers count: mental disorders in America: A summary of statistics describing the prevalence of mental disorders in America. NIMH Report (www.nimh.nih.gov/publicat/numbers.cfm) 2001;24:185–195.
6. Mergl R, Seidscheck I, Allgaier AK, Moller HJ, Hegerl U, Henkel V. Depressive, anxiety, and somatoform disorders in primary care: prevalence and recognition. Depress Anxiety; 2006.
7. Sherbourne CD, Jackson CA, Meredith LS, Camp P, Wells KB. Prevalence of comorbid anxiety disorders in primary care outpatients. Arch Fam Med 1996;5:27–34.
8. Barsky AJ, Cleary PD, Coeytaux RR, Ruskin JN. Psychiatric disorders in medical outpatients complaining of palpitations. J Gen Intern Med 1994;9:306–313.
9. Dahl AA, Haaland CF, Mykletun A, Bremnes R, Dahl O, Klepp O, et al. Study of anxiety disorder and depression in long-term survivors of testicular cancer. J Clin Oncol 2005;23:2389–2395.
10. Katon W, Von Korff M, Lin E, Walker E, Simon G.E, Bush T, et al. Collaborative management to achieve treatment guidelines. Impact on depression in primary care. J Am Med Assoc 1995;273:1026–1031.
11. Stein MB, Asmundson GJ, Ireland D, Walker JR. Panic disorder in patients attending a clinic for vestibular disorders. Am J Psychiatry 1994;151:1697–1700.
12. Walker EA, Roy-Byrne PP, Katon WJ, Li L, Amos D, Jiranek G. Psychiatric illness and irritable bowel syndrome: a comparison with inflammatory bowel disease. Am J Psychiatry 1990; 147:1656–1661.
13. DuPont RL, Rice DP, Miller LS, Shiraki SS, Rowland CR, Harwood HJ. Economic costs of anxiety disorders. Anxiety 1996;2:167–172.
14. Balon R. Mood, anxiety, and physical illness: body and mind, or mind and body? Depress Anxiety 2006;23:377–387.
15. Barger SD, Sydeman SJ. Does generalized anxiety disorder predict coronary heart disease risk factors independently of major depressive disorder? J Affect Disord 2005;88:87–91.
16. Bittner A, Goodwin RD, Wittchen HU, Beesdo K, Hofler M, Lieb R. What characteristics of primary anxiety disorders predict subsequent major depressive disorder? J Clin Psychiatry 2004;65:618–626.
17. Fleishman JA, Cohen JW, Manning WG, Kosinski M. Using the SF-12 health status measure to improve predictions of medical expenditures. Med Care 2006;44:I54–I63.
18. Hadjistavropoulos HD, Asmundson GJ, Kowalyk KM. Measures of anxiety: is there a difference in their ability to predict functioning at three-month follow-up among pain patients? Eur J Pain 2004;8:1–11.
19. Hornbrook MC, Goodman MJ. Assessing relative health plan risk with the RAND-36 health survey. Inquiry 1995;32:56–74.
20. Nease Jr DE, Volk RJ, Cass AR. Does the severity of mood and anxiety symptoms predict health care utilization? J Fam Pract 1999;48:769–777.
21. Spielberger CD. Manual for the State-Trait Anxiety Inventory (STAI). PaloAlto, CA: Consulting Psychologists Press; 1983.
22. Beck AT, Steer RA. Beck Anxiety Inventory. BAI. San Antonio: The Psychological Cooperation; 1993.
23. Hermann Ch, Buss U, Snaith RP. Hospital Anxiety and Depression Scale: HADS. Bern: Verlag Hans Huber; 1995.
24. Zung WW. A rating instrument for anxiety disorders. Psychosomatics 1971;12:371–379.
25. Nunnally JC, Bernstein C. Psychometric theory. (2nd ed.) New York: McGraw-Hill; 1994.
26. Embretson S, Reise SP. Item Response Theory for Psychologists. Mahwah: Lawrence Earlbaum Associates, Publishers; 2000.
27. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, California: Sage Publications, Inc.; 1991.
28. van der Linden WJ. Hambleton RK. Handbook of modern item response theory. Berlin: Springer; 1997.
29. Gershon R.C. Computer adaptive testing. J Appl Meas 2005; 6:109–127.
30. Meijer RR, Nering ML. Computerized adaptive testing. Overview and introduction. Appl Psychol Meas 1999;23: 187–194.
31. Wainer H, Dorans NJ, Eignor D, et al. Computer-adaptive testing: a primer. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
32. Bjorner JB, Kosinski M, Ware Jr JE. Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. Qual Life Res 2003;12: 981–1002.
33. Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential functioning of the Beck depression inventory in late-life patients: use of item response theory. Psychol Aging 2002;17:379–391.

34. Hammond SM. An IRT investigation of the validity of non-patient analogue research using the Beck Depression Inventory. Eur J Psychol Assess 1995;11:14–20.
35. Santor DA, Ramsay JO, Zuroff DC. Nonparametric item analysis of the Beck Depression Inventory: evaluating gender item bias and response option weights. Psychol Assess 1994;6: 255–270.
36. Cooke DJ, Michie C. An item response theory analysis of the Hare Psychopathy Checklist—Revised. Psychol Assess 1997; 9:3–14.
37. Walter O, Becker J, Fliege H, Bjorner JB, Kosinski M, Klapp BF, et al. Developmental steps for a computer adaptive test for anxiety (A-CAT). Diagnostica 2005;51:88–100.
38. Cook KF, Taylor PW, Dodd BG, Teal CR, McHorney CA. Evidence-based practice for equating health status items: sample size and IRT model. J Appl Meas 2007;8:175–189.
39. Dorans NJ. Linking scores from multiple health outcome instruments. Qual Life Res 2007;16:85–94.
40. Jones RN, Fonda SJ. Use of an IRT-based latent variable model to link different forms of the CES-D from the Health and Retirement Study. Soc Psychiatry Psychiatr Epidemiol 2004;39: 828–835.
41. Orlando M, Sherbourne CD, Thissen D. Summed-score linking using item response theory: application to depression measurement. Psychol Assess 2000;12:354–359.
42. Reise SP, Henson JM. Computerization and adaptive administration of the NEO PI-R. Assessment 2000;7:347–364.
43. Zhu W. Test equating: what, why, how? Res Q Exerc Sport 1998;69:11–23.
44. Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton Depression Rating Scale measure? J Psychiatr Res 1993;27:259–273.
45. Rose M, Bjorner J, Becker J, Fries J, Ware JE. Evaluation of a preliminary physical function item bank supports the expected advantages of the patient-reported outcomes measurement information system (PROMIS). J Clinical Epidemiol 2008;61: 17–33.
46. Ware Jr JE, Bjorner JB, Kosinski M. Dynamic health assessment: the search for more practical and more precise outcome measures. Qual Life Newsl 1999;January:11–13.
47. Butcher JN, Keller LS, Bacon SF. Current developments and future directions in computerized personality assessment. J Consult Clinical Psychol 1985;53:803–815.
48. Becker J. Computergestuetztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT) [Computerized Adaptive Testing (CAT) of anxiety based on Item Response Theory (IRT)]. Berlin, Germany: Free University of Berlin; 2004.
49. Becker J, Walter OB, Fliege H, Rose M, Klapp BF. A computer adaptive test for measuring anxiety. Psychother Psychosom Med Psychol 2004;53:99.
50. Dilling H, Mombour W, Schmidt MH. Internationale Klassifikation psychischer Störungen. ICD-10 Kapitel V (F). Klinisch-diagnostische Leitlinien (3rd ed.) Bern: Huber; 1999.
51. Hoerhold M, Klapp BF. Testung der Invarianz und der Hierarchie eines mehrdimensionalen Stimmungsmodells auf der Basis von Zweipunkterhebungen an Patienten- und Studentenstichproben. Z Med Psychol 1993;2:27–35.
52. Ware Jr JE, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38:II73–II82.
53. Ware Jr JE, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlof CG, et al Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res 2003;12:935–952.
54. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982;12: 261–280.
55. Margraf J, Ehlers A. Beck Angst Inventar (Beck Anxiety Inventory, original version by A.T. Beck and R.A. Stern). Göttingen: Hogrefe; 2007.
56. Laux L, Glanzmann P, Schaffner P, Spielberger CD. STAI State-Trait-Angstinventar (Spielberger, C.D, Gorsuch, R.L, Lushene, R.E, 1970). Test manual. Weinheim: Beltz; 1981.
57. Rose M, Walter OB, Fliege H, Becker J, Hess V. 7 years of experience using Personal Digital Assistants (PDA) for psychometric diagnostics in 6000 inpatients and polyclinic patients. In: Bludau HB, Koop A, editors. Mobile computing in medicine. GI-edition lecture notes in informatics, P-15. Köllen: Verlag; 2002:35–44.
58. Gardner W, Kelleher KJ, Pajer KA. Multidimensional adaptive testing for mental health problems in primary care. Med Care 2002;40:812–823.
59. Kobak KA, Greist JH, Jefferson JW, Katzelnick DJ. Computer-administered clinical rating scales. A rev. Psychopharmacol (Berl) 1996;127:291–301.
60. Allenby A, Matthews J, Beresford J, McLachlan SA. The application of computer touch-screen technology in screening for psychosocial distress in an ambulatory oncology setting. J Cancer Care (Engl) 2002;11:245–253.
61. Carlson LE, Speca M, Hagen N, Taenzer P. Computerized quality-of-life screening in a cancer pain clinic. J Palliat Care 2001;17:46–52.
62. Sigle J, Porzsolt F. Practical aspects of quality-of-life measurement: design and feasibility study of the quality-of-life recorder and the standardized measurement of quality of life in an outpatient clinic. Cancer Treat Rev 1996;22:75–89.
63. Simms LJ, Clark LA. Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). Psychol Assess 2005;17:28–43.
64. Velikova G, Brown JM, Smith AB, Selby PJ. Computer-based quality of life questionnaires may contribute to doctor-patient interactions in oncology. Br J Cancer 2002;86:51–59.
65. Wilkie DJ, Judge MK, Berry DL, Dell J, Zong S, Gilespie R. Usability of a computerized PAINReportIt in the general public with pain and people with cancer pain. J Pain Symptom Manage 2003;25:213–224.
66. Wilson AS, Kitas GD, Carruthers DM, Reay C, Skan J, Harris S, et al. Computerized information-gathering in specialist rheumatology clinics: an initial evaluation of an electronic version of the Short Form 36. Rheumatology (Oxford) 2002;41:268–273.
67. Hornke LF. Benefits form computerized adaptive testins as seen in simulation studies. Eur J Psychol Assess 1999;15:91–98.
68. Ware Jr JE, Gandek B, Sinclair SJ, Bjorner JB. Item response theory and computerized adaptive testing: implications for outcomes measurement in rehabilitation. Rehabil Psychol 2005;50:71–78.
69. Bayliss MS, Dewey JE, Dunlap I, Batenhorst AS, Cady R, Diamond ML, et al. A study of the feasibility of internet administration of a computerized health survey: The Headache Impact Test (HITTM). Qual Life Res 2003;12:953–961.
70. Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, et al. Computerized adaptive measurement of depression: a simulation study. BMC Psychiatry 2004;4:1–11.

71. Kosinski M, Bjorner JB, Ware Jr JE, Sullivan E, Straus WL. An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. J Clin Epidemiol 2006;59:715–723.
72. Schwartz C, Welch G, Santiago-Kelley P, Bode R, Sun X. Computerized adaptive testing of diabetes impact: a feasibility study of Hispanics and non-Hispanics in an active clinic population. Qual Life Res 2006;15:1503–1518.
73. Blankstein KR. The sensation seeker and anxiety reactivity: relationships between the sensation-seeking scales and the activity preference questionnaire. J Clin Psychol 1975;31: 677–681.
74. Enns MW, Cox BJ, Parker JD, Guertin JE. Confirmatory factor analysis of the Beck Anxiety and Depression Inventories in patients with major depression. J Affect Dis 1998;47:195–200.
75. Cox BJ, Cohen E, Direnfeld DM, Swinson RP. Does the Beck Anxiety Inventory measure anything beyond panic attack symptoms? Behav Res Ther 1996;34:949–954.
76. Creamer M, Foran J, Bell R. The Beck Anxiety Inventory in a non-clinical sample. Behav Res Ther 1995;33:477–485.
77. Forbey JD, Ben Porath YS. Computerized adaptive personality testing: a review and illustration with the MMPI-2 Computerized Adaptive Version. Psychol Assess 2007;19:14–24.
78. Brown TA, Campbell LA, Lehman CL, Grisham JR, Mancill RB. Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. J Abnorm Psychol 2001;110:585–599.
79. Mineka S, Watson D, Clark LA. Comorbidity of anxiety and unipolar mood disorders. Ann Rev Psychol 1998;49:377–412.
80. Roy-Byrne PP, Stang P, Wittchen HU, Ustun B, Walters EE, Kessler RC. Lifetime panic-depression comorbidity in the National Comorbidity Survey: association with symptoms, impairment, course and helpseeking. Br J Psychiatry 2000;176: 229–235.
81. Barlow DH, Chorpita BF, Turovsky J. Fear, panic, anxiety and disorders of emotion. Nebr Symp Motiv 1996;43:251–328.
82. Cannon MF, Weems CF. Do anxiety and depression cluster into distinct groups? A test of tripartite model predictions in a community sample of youth. Depress Anxiety 2006;23:453–460.
83. Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. J Abnorm Psychol 1995;104:15–25.
84. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005;23:S53–S57.
85. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res 2005;14:2277–2291.
86. Kocalevent RD, Walter O, Becker J, Fliege H, Klapp BF, Bjorner JB, et al. Stress-CAT—a computer adaptive test for the measurement of stress experience (Abstract). Psychother Psychosom Med Psychol 2005;55:136.
87. Kubinger KD, Wurst E. Adaptives Intelligenz Diagnostikum 2. AID2. Göttingen: Beltz; 2000.
88. Handel RW, Ben-Porath YS, Watt M. Computerized adaptive assessment with the MMPI-2 in a clinical setting. Psychol Assess 1999;11:369–380.
89. Roper BL, Ben Porath YS, Butcher JN. Comparability of computerized adaptive and conventional testing with the MMPI-2. J Pers Assess 1991;57:278–290.
90. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003;12:485–501.
91. Bode RK, Cella D, Lai JS, Heinemann AW. Developing an initial physical function item bank from existing sources. J Appl Meas 2003;4:124–136.
92. Siebens H, Andres PL, Pengsheng N, Coster WJ, Haley SM. Measuring physical function in patients with complex medical and postsurgical conditions: a computer adaptive approach. Am J Phys Med Rehabil 2005;84:741–748.
93. Haley SM, Raczek AE, Coster WJ, Dumas HM, Fragala-Pinkham MA. Assessing mobility in children using a computer adaptive testing version of the pediatric evaluation of disability inventory. Arch Phys Med Rehabil 2005;86:932–939.
94. Hoyer J, Becker ES, Neumer S, Soeder U, Margraf J. Screening for anxiety in an epidemiological sample: predictive accuracy of questionnaires. J Anxiety Disord 2002;16:113–134.
95. Espallargues M, Valderas JM, Alonso J. Provision of feedback on perceived health status to health care professionals: a systematic review of its impact. Med Care 2000;38:175–186.
96. Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? Soc Sci Med 2005;60:833–843.
97. Brown MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. Testing structural equation models. Newbury Park, CL: Sage; 1993:136–162.
98. MacCallum RC, Brwon MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1:130–149.
99. Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof. Sep 2005;28(3)264–282
100. Edelen M, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research 2007;16(1)5–18
101. Fliege H, Becker J, Walter OB, Rose M, Bjorner JB, Klapp BF. (2008). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. [in review 2008].