Computerized adaptive testing in Dutch mental health care: a new tool to assess depression and anxiety

Slok-Flens, G.

Chapter 6

Construct validity, Responsiveness, and Utility of Change Indicators of the Dutch-Flemish PROMIS Item banks for Depression and Anxiety Administered as Computerized Adaptive Test (CAT): A Comparison with the Brief Symptom Inventory (BSI).

Published as:

Flens, G., Terwee, C. B., Smits, N., Williams, G., Spinhoven, P., Roorda, L. D., & de Beurs, E. (2022). Construct validity, responsiveness, and utility of change indicators of the Dutch-Flemish PROMIS item banks for depression and anxiety administered as computerized adaptive test (CAT): A comparison with the Brief Symptom Inventory (BSI). *Psychological Assessment, 34*(1), 58-69.

6.1 Abstract

We evaluated construct validity, responsiveness, and utility of change indicators of the Dutch-Flemish PROMIS adult v1.0 item banks for Depression and Anxiety administered as computerized adaptive test (CAT). Specifically, the CATs were compared to the Brief Symptom Inventory (BSI) using pretest and retest data of adult patients treated for common mental disorders (N = 400; median pretest to retest interval = 215 days). Construct validity was evaluated with Pearson's correlations and Cohen's *d*s; responsiveness with Pearson's correlations and pre-post effect sizes (*ES*); utility of change indicators with kappa coefficients and percentages of (dis)agreement. The results showed that the PROMIS CATs measure similar constructs as matching BSI scales. Under the assumption of measuring similar constructs, the CAT and BSI Depression scales were similarly responsive. For the Anxiety scales, we found a higher responsiveness for CAT (*ES* = 0.64) compared to the BSI (*ES* = 0.50). Finally, both CATs categorized the change scores of more patients as changed compared to matching BSI scales, indicating that the PROMIS CATs may be more able to detect actual change than the BSI. Based on these findings, the PROMIS CATs may be considered a modest improvement over matching BSI scales as tools for reviewing treatment progress with patients. We discuss several additional differences between the PROMIS CATs and the BSI to help test users choose instruments. These differences include the adopted measurement theory (Item Response Theory vs. Classical Test Theory), the mode of administration (CAT vs. fixed items), and the area of application (universal vs. predominantly clinical).

Keywords: clinical assessment, depression, anxiety, PROMIS CAT, psychometric properties

6.2 Background

In Dutch health care, computerized adaptive tests (CATs) are gradually being implemented to evaluate self-reported health in clinical subjects (e.g., depression, physical function, and ability to participate in social roles and activities; Terwee et al., 2014). A CAT is a computer-based test in which items are administered from an item bank (i.e., a set of items that measure a specific construct) according to the answers to previous selected items, and that terminates when a stopping rule is met (e.g., a specific measurement precision). As a result, patient burden can be reduced with a shorter measurement and a negligible loss of precision (Fliege et al., 2005).

The first item banks that were psychometrically evaluated for CAT administration in Dutch *mental* health care were the Patient-Reported Outcomes Measurement Information System (PROMIS®) adult v1.0 item banks for Depression and Anxiety. In previous studies, these item banks were translated into Dutch-Flemish (DF; Terwee et al., 2014) and psychometrically evaluated for cross-sectional (Flens et al., 2017, 2019) and longitudinal applications (Flens et al., 2021). The cross-sectional studies showed that both item banks have good quality item parameters according to the PROMIS standards (Reeve et al., 2007). Moreover, post hoc CAT simulations showed that both item banks, when administered adaptively, can be highly precise as well as efficient in both the general population and clinical ambulatory populations at the start of treatment. In addition, the longitudinal study showed that, using tentative rules of thumb, the Depression and Anxiety item banks were sufficiently invariant over time in clinical samples with mood and anxiety disorders, respectively. In other words, the item banks appear to provide (change) scores that reflect single depression and anxiety constructs.

The results of these earlier studies indicate that the DF PROMIS adult v1.0 item banks for Depression and Anxiety have adequate psychometric properties for both cross-sectional and longitudinal applications. However, the item banks still need to be validated with actual CAT administrations, and compared to an established Dutch legacy instrument before introducing them in routine assessment of clinical subjects. After all, we want to ensure that the psychometric properties of the PROMIS CATs are at least as good as those of legacy instruments to convince users that changing instruments results in similar (and preferably even better) assessment of patients.

Psychometric properties that demand additional attention are construct validity and responsiveness (Maruyama & Ryan, 2014; Mokkink et al., 2010; Pilkonis et al., 2014). Furthermore, the utility of reliability-based indicators of clinical significant change need to be evaluated to facilitate the use of the PROMIS CATs in clinical practice (Jacobson & Truax, 1991). These aspects are seen as relevant because they reflect an instrument's ability to aid professionals in planning treatments, evaluating therapeutic interventions, and anticipating and planning timely termination (de Beurs et al., 2018). Furthermore, regular or continuous monitoring of progress with appropriate and psychometrically sound instruments may help to prevent treatment failure (Lambert, 2010).

In previous studies, using clinical samples, it was demonstrated that the Unites States (US) PROMIS instruments for Depression and Anxiety (i.e., CATs and short-forms) measure similar constructs as legacy instruments, and are similarly responsive (Kroenke et al., 2019; Pilkonis et al., 2014). These results were shown for the PROMIS Depression instruments compared to the legacy instruments Center for Epidemiological Studies Depression scale (CESD) and Patient Health Questionnaire (PHQ-9), and for the PROMIS Anxiety instruments compared to the legacy instruments Generalized Anxiety Disorder (GAD-7), Symptom Checklist (SCL), Posttraumatic Stress disorder checklist (PCL), Short Form (SF)-36, and SF-12 Mental Component Summary (MCS). We therefore expect that the DF PROMIS CATs for Depression and Anxiety also measure similar constructs as Dutch legacy instruments, and are at least as responsive. In addition, Pilkonis et al. (2014) showed that the US PROMIS CAT for Depression measures more reliably than the legacy instruments CESD and PHQ-9, probably because CAT ensures that each administration meets the minimally required measurement precision, by which the number of administered items is allowed to vary among respondents. The legacy instruments, on the other hand, fix the number of items, by which the measurement precision will vary among respondents. Based on these measurement properties, we expect that reliability-based indicators of clinical significant change categorize more patients as actually changed for the DF PROMIS CATs compared to fixed-item legacy instruments.

This study was the first in the Netherlands in which PROMIS CATs were administered. We aimed to assess construct validity, responsiveness, and utility of change indicators of the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CAT in a clinical sample. Specifically, the PROMIS CATs were compared to the nine subscales of the Dutch Brief Symptom Inventory (BSI; de Beurs & Zitman, 2005; Derogatis et al., 1973) using pretest and retest data. We chose the BSI as legacy instrument because two of its subscales aim to measure the same constructs as the CATs; it is often used as outcome measure in routine assessment of patients internationally and in the Netherlands; and it has been claimed to have adequate psychometric properties for both cross-sectional and longitudinal applications (Carlier et al., 2017; de Beurs & Zitman, 2005; van Noorden et al., 2010). More specifically, it has been demonstrated that the BSI is sufficiently reliable, valid and responsive compared to a large number of legacy instruments. These include the Symptom Questionnaire-48 (SQ-48), the Outcome Questionnaire-45 (OQ-45), and several disorder-specific instruments (e.g., the Montgomery Äsberg Depression Rating Scale [MADRS], Beck Depression Inventory [BDI-II], Padua Inventory [PI], Yale Brown Obsessive Compulsive Scale [Y-BOCS], and Panic Disorder Severity Scale [PDSS].

6.3 Methods

6.3.1 Participants

Data were collected between September 2017 and June 2019 in a clinical population of adult patients who started outpatient treatment for common mental disorders. Patients were invited by the Dimence Group, which is a large mental health institute offering inpatient and outpatient treatment in the eastern part of the Netherlands. The patient's diagnosis (*Diagnostic and*

Statistical Manual of Mental Disorders, 5th ed.; *DSM-5*; American Psychiatric Association, 2013) was assessed by a therapist in a clinical face-to-face assessment (i.e., the intake of treatment).

This study has not been submitted to a research ethics committee because, according to Dutch law, data collected as part of clinical practice may be used in anonymized form for scientific research (de Beurs et al., 2011). Consequently, all data were coded before they were released to the first author for analysis, and could not be traced back to a person by the authors. This was approved by the privacy and information security officer of the Dimence Group. In addition, patients were informed upon their referral and registration for treatment that their data might be used for research, and that an opt-out procedure was available if they did not consent to this. Data from patients with an opt-out registration were not released to the first author.

In accordance with similar studies, we aimed to include at least 200 patients (Pilkonis et al., 2014; Schalet et al., 2016). A patient was included when (a) a pretest and retest score were available for all measures to perform the analyses in a straightforward fashion (i.e., without missing cases), (b) the measures were completed on the same day for both pretest and retest to establish a set of instruments that was administered under similar conditions as much as possible, (c) the retest was administered at least one month after the pretest to increase the possibility that at least some change had occurred between measurements, and (d) the retest was administered after the first treatment session to ensure that at least some treatment was provided.

6.3.2 Measures

The measures were part of a larger battery of instruments to be completed by the patients, and consisted of the DF PROMIS adult v1.0 item banks for Depression (Flens et al., 2017) and Anxiety (Flens et al., 2019) administered as CAT, and the Dutch BSI (de Beurs & Zitman, 2005). For each patient, the measures were administered digitally through an automated process. In this process, the PROMIS CATs were assigned in alternating order for both pretest and retest: the CAT Anxiety was administered first at even weeks, the CAT Depression was administered first at odd weeks. The BSI was always administered directly after the PROMIS CATs. According to Dimence Group's policy, the invitation for the pretest was sent before or during the intake session.

6.3.2.1 PROMIS CATs

The content of the DF PROMIS adult v1.0 item banks for Depression and Anxiety item banks reflects a wide range of depression and anxiety symptoms, problems, or negative affective states (e.g., Depression item bank, EDDEP04 *I felt worthless*; Anxiety item bank, EDANX01 *I felt fearful*). Respondents were asked by computer to indicate on a 5-point scale how frequently they experienced the symptoms, problems or negative states in the past 7 days (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*), a higher score indicating more severe depression or anxiety.

For the CAT item selection algorithm, we followed the recommendations of other studies (Flens et al., 2017, 2019), using Fisher's information function calculated with the DF item parameters. The initial item was selected as the item with the greatest Fisher's information

at the value of the estimated mean (*M*) of the latent trait for the Dutch general population. For the CAT Depression this item was EDDEP36 *I felt unhappy*; for the CAT Anxiety EDANX54 *I felt tense* was selected. After each item, the maximum likelihood estimate (MLE) of the test taker's latent trait was calculated. Each sequential item was then selected as the item with the greatest Fisher's information at the value of the MLE. The CAT was terminated when either the measurement precision fell below a predefined threshold or the upper limit of administered items was reached. The measurement precision threshold was set to a *SE*(θ) below .22, with the *SE*(θ) approximated as the reciprocal of the square root of the information function. The threshold of .22 was selected to be comparable to a marginal reliability of .95 (Green et al., 1984), which is considered a high standard for the precision of assessments that provide scores to individuals (Bernstein & Nunnally, 1994). The upper limit of administered items was set to nine for the CAT Depression, and to 12 for the CAT Anxiety (Flens et al., 2019).

According to PROMIS convention, we used the item responses, the US item parameters, and the expected a posteriori (EAP) estimator to calculate PROMIS T-scores and their accompanying measurement precision through the HealthMeasures provided Scoring Service, powered by Assessment Center (www.assessmentcenter.net/ac_scoringservice). PROMIS T-scores are represented on a scale with a *M* of 50 and a standard deviation (*SD*) of 10 in the general US population, a higher score indicating more severe depression or anxiety.

6.3.2.2 BSI

The BSI is a multidimensional self-report instrument that evaluates the severity of symptoms of psychopathology. The instrument consists of an 53-item overall scale, divided into nine subscales (i.e., Depression, 6 items; Anxiety, 6 items; Somatization, 7 items; Obsessive-Compulsive, 6 items; Interpersonal Sensitivity, 4 items; Hostility, 5 items; Phobic Anxiety, 5 items; Paranoid Ideation, 5 items; Psychoticism; 5 items) and four remaining items. For this study, we used the Depression subscale (e.g., item 18 *feeling no interest in things*) and the Anxiety subscale (e.g., item 38 *feeling tense or keyed up*) to evaluate the CATs on their relation with scales measuring matching constructs. The other subscales were used to evaluate the CATs on their relation with scales measuring other constructs. For all subscales, respondents were asked by computer to indicate on a 5-point scale to what extent they were bothered by the symptoms, problems or negative states in the past 7 days (0 = *not at all*, 1 = *a little bit*, 2 = *moderately*, 3 = *quite a bit*, and 4 = *extremely*). Average scores were calculated for each subscale (ranging from 0 – 4), a higher score meaning more distress.

6.3.3 Statistical analyses

We performed analyses to report on descriptive statistics, construct validity, responsiveness, and utility of change indicators. A hypothesis was formulated for each analysis to compare the instruments. As rule of thumb, we considered a psychometric property as sufficiently supported when at least 75% of the hypotheses were confirmed (Prinsen et al., 2018). For indicators of change between pretest and retest scores, we did not correct for pretest severity (O'Connell et al., 2017). All statistical analyses were performed in the statistical environment R (R Core Team, 2018).

6.3.3.1 Descriptive statistics

Based on the inclusion criteria, we assessed the gender- and age distribution of the study sample. Furthermore, we evaluated whether the composition of the study sample was representative for the mental health provider that collected the data. To accomplish this, it was assessed whether the included patients were similar to the nonincluded patients regarding the distribution of gender, age, and pretest score. For gender, we investigated Pearson's residuals as measure of effect size, following the suggestion of 2.00 as cutoff value for indicating a substantial difference between the observed respondents and the expected number of respondents under the model (Agresti & Kateri, 2011). For age and pretest score, we investigated Cohen's *d* as measure of effect size (i.e., the *M* difference divided by the pooled *SD*), following the guideline proposed by Cohen (1988) to interpret the size of the effect: 0.20 = small effect, 0.50 = medium effect, and 0.80 = large effect.

In addition, we assessed the mean number of administered items for both pretest and retest of the Depression and Anxiety scales. Furthermore, we evaluated the variation in pretest to retest interval by calculating quantiles of the days between the tests.

6.3.3.2 Construct validity

A classic definition of construct validity is the degree to which a test measures the concept it is supposed to measure (Cook & Campbell, 1979). We investigated this psychometric property by collecting multiple sources of empirical evidence commonly claimed as indicative of validity (Newton & Shaw, 2014).

First, we studied convergent and divergent validity by evaluating whether the measured constructs of the PROMIS CATs are related to those of matching BSI scales, and unrelated to those of other BSI scales (Cook & Campbell, 1959). For convergent validity, it was hypothesized that Pearson's correlation coefficients between the CATs and matching BSI scales were above 0.50 (Prinsen et al., 2018) for both pretest (*Hypothesis 1*) and retest (*Hypothesis 2*). For divergent validity, it was hypothesized for both pretest (*Hypothesis 3*) and retest (*Hypothesis 4*) that Pearson's correlation coefficients between the CATs and other BSI scales were at least 0.10 points lower than those between the CATs and matching BSI scales (Prinsen et al., 2018).

Next, we studied concurrent validity by evaluating whether the PROMIS CATs are at least as able as matching BSI scales to distinguish between distinct groups based on the patient's primary diagnosis (i.e., the condition that causes the patient the most problems or discomfort, as assessed at the intake of treatment; American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1974). Consequently, the study sample was divided into patients with and without a primary depression diagnosis to compare the Depression scales, and into patients with and without a primary anxiety diagnosis to compare the Anxiety scales. We then compared Cohen's *d* measure of effect size with a 95% confidence interval (*CI*) between the CATs and matching BSI scales (Hedges & Olkin, 2014). Cohen's *d* was calculated as the *M* score difference between patients with and without a primary diagnosis divided by the pooled *SD* of these subsamples. We suggest that a difference in *d*-values of at most 0.10 points indicates sufficient similarity in the ability

to discriminate between patients with and without a specific disorder. Consequently, it was hypothesized that the *d*-values of the CATs were at most 0.10 points lower than those of matching BSI scales (*Hypothesis 5*). This was evaluated for the pretest only because the primary diagnosis was assessed around this test.

Finally, we studied stability by evaluating whether the pretest to retest associations of the PROMIS CATs are sufficiently similar to those of matching BSI scales (Drenth & Sijtsma, 2005). To study stability, we suggest that a difference in Pearson's pretest to retest correlation coefficients of at most 0.10 points indicates sufficient similarity in stability. Consequently, it was hypothesized that the pretest to retest correlation coefficients of the CATs differed at most 0.10 points from those of matching BSI scales (*Hypothesis 6*).

6.3.3.3 Responsiveness

Responsiveness is defined by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) as the ability of an instrument to detect change over time in the construct to be measured (Mokkink et al., 2010). To study this psychometric property, we evaluated whether the change scores of the PROMIS CATs are related to those of matching BSI scales, and unrelated to those of other BSI scales. It was hypothesized that Pearson's correlation coefficients between the change scores of the CATs and matching BSI scales were above 0.50 (*Hypothesis 7*; Prinsen et al., 2018). Furthermore, it was hypothesized that Pearson's correlation coefficients between the change scores of the CATs and other BSI scales were at least 0.10 points lower than those between the CATs and matching BSI scales (*Hypothesis 8*; Prinsen et al., 2018).

In addition, pre-post effect sizes (*ES*) with a 95% *CI* were used to evaluate whether the PROMIS CATs are at least as responsive as matching BSI scales (Seidel et al., 2014). *ES* is calculated as the *M* change score of a scale divided by the *SD* of that scale's pretest scores. We suggest that a difference in *ES* values of at most 0.10 points indicates sufficient similarity in responsiveness. Consequently, it was hypothesized that the *ES* values of the CATs were at most 0.10 points lower than those of matching BSI scales (*Hypothesis 9*).

6.3.3.4 Utility of change indicators

To evaluate whether patients improve or deteriorate, often-used indicators are reliable change and clinically significant change (CSC; Jacobson & Truax, 1991). Reliable change is defined as a change in scores that may not have occurred due to random measurement error alone. *CSC* is defined as a change from a clinical population score to a general population score. We combined reliable change and *CSC* to evaluate whether the PROMIS CATs categorize more patients as actually changed than matching BSI scales (de Beurs et al., 2019).

Reliable change was evaluated with the *Z*-test for the CATs (Brouwer et al., 2013) and with the reliable change index (*RCI*) for the BSI scales (Jacobson & Truax, 1991). Different methods were used because the CATs and the BSI assume different measurement theories (i.e., item response theory [IRT] and classical test theory [CTT], respectively). To assess reliable change, we used the SEs of a patient's pretest and retest for the *Z*-test, and the test-retest reliability as determined by de Beurs and Zitman (2005) for the *RCI*. A *Z*- or *RCI* value larger

than 1.96 reflects with a 95% *CI* that the change in pretest to retest scores may not have occurred due to random measurement error alone (Brouwer et al., 2013).

The cutoff for *CSC* was calculated as the point halfway the general- and clinical population. To determine this value for each Depression and Anxiety scale, we used the samples of previous psychometric studies for the general population (CAT Depression, Flens et al., 2017; CAT Anxiety, Flens et al., 2019; BSI, de Beurs & Zitman, 2005), and the pretest sample of the current study for the clinical population. Subsequently, we used the cutoff values for both *CSC* and reliable change to categorize the patients into four groups: recovered (CATs, $Z \geq 1.96$; BSI, $RCI \geq 1.96$; pretest score > *CSC*; retest score ≤ *CSC*), improved (CATs, $Z \geq 1.96$; BSI, $RCI \geq 1.96$), unchanged (CATs, $-1.96 \leq Z < 1.96$; BSI, $-1.96 \leq RCI < 1.96$), and deteriorated (CATs, $Z < -1.96$; BSI, $RCI < -1.96$).

We used the modified Fleiss kappa statistic for ordinal variables (i.e., the *s** statistic) with linear weights and a 95% *CI* (Marasini et al., 2016) as well as the percentage of agreement to assess whether the PROMIS CATs showed a substantial disagreement with matching BSI scales in categorizations, and, if so, whether the CATs categorized the change scores of less patients as unchanged. We considered this to be the case when three criteria where met: the *s** statistic was smaller than 0.60 (McHugh, 2012), the percentage of agreement was smaller than 80% (McHugh, 2012), and the percentage of patients that were categorized as unchanged was smaller for the CATs than for matching BSI scales (*Hypothesis 10*).

The data of this study are not publicly available because they were used under license from the Dimence Group. However, the data can be made available from the first author upon reasonable request and with permission of the Dimence Group. The study analysis code can be requested from the first author. This study was not preregistered.

6.4 Results

6.4.1 Descriptive statistics

The eligible sample (i.e., the patients that were invited for the pretest and retest) consisted of 549 respondents. Of these respondents, 544 responded to the pretest (response rate = 99.1%) and 504 also responded to the retest (response rate = 91.8%). Furthermore, we excluded 104 respondents for not meeting the remaining inclusion criteria. Consequently, our final sample consisted of $N = 400$ (response rate = 72.9%; 64.0% female; age $M = 37.4$ years, $SD = 12.2$, range 18–66). For this sample, 46% of the patients had a mood disorder as the primary reason for seeking treatment, 39% had an anxiety disorder, and 15% had another disorder (e.g., attention deficit disorder, somatoform disorder, personality disorder). In addition, the pretest and retest did not include missing item responses. Consequently, the analyses were performed in a straightforward fashion.

Next, the comparison between the included and nonincluded patients showed that Pearson's residuals were all below 2.00 for gender, and Cohen's *d*s were all below 0.20 for age

and pretest score. These results indicate that the included patients were sufficiently similar to the nonincluded patients for the variables gender, age, and pretest score.

Finally, the *M* (*SD*) number of administered items was 6.7 (1.0) for the CAT Depression pretest (7% responded to all 9 items), 6.6 (1.0) for the CAT Depression retest (9% responded to all 9 items), 8.7 (1.1) for the CAT Anxiety pretest (3% responded to all 12 items), and 8.5 (1.1) for the CAT Anxiety retest (3% responded to all 12 items). For the BSI, all patients responded to the six fixed items of the Depression scale and the six fixed items of the Anxiety scale. In addition, the median of the pretest to retest interval was 215 days (range = 32–505, interquartile range = 145–281), indicating a substantial variation in intervals.

6.4.2 Construct validity

Table 6.1 displays Pearson's correlation coefficients between the PROMIS CATs and BSI subscales for the pretest and retest. In support of *Hypothesis 1* (pretest) and *Hypothesis* 2 (retest), the correlation coefficients between the CATs and matching BSI scales were above .50 (Depression, pretest r = .83, retest r = .87; Anxiety, pretest r = .76, retest r = .81). Furthermore, in support of *Hypothesis 3* (pretest) and *Hypothesis 4* (retest), the correlation coefficients between the CATs and other BSI scales were at least 0.10 points below those between the CATs and matching BSI scales. Note that the CATs correlated lower with most of the other BSI scales than the BSI Depression and Anxiety scales did.

In support of *Hypothesis 5*, the *d*-values between patients with and without a specific primary diagnosis (i.e., depression or anxiety) were not more than 0.10 points lower for the PROMIS CATs relative to those between matching BSI scales. The comparison between the depression (n = 184; CAT M = 65.7, SD = 6.4; BSI M = 2.21, SD = 0.98) and not-depression subsamples (n = 216; CAT M = 63.5, SD = 7.0; BSI M = 1.95, SD = 0.96) resulted in Cohen's d = 0.33, 95% *CI* [0.13, 0.53] for the CAT Depression, and d = 0.26, 95% *CI* [0.06, 0.46] for the BSI Depression scale. The comparison between the anxiety (n = 157; CAT M = 68.1, SD = 6.0; BSI M = 1.97, SD = 0.93) and not-anxiety subsamples (n = 243; CAT M = 66.6, SD = 6.6; BSI M = 1.66, SD = 0.91) resulted in d = 0.24, 95% *CI* [0.04, 0.44] for the CAT Anxiety and d = 0.33, 95% *CI* [0.13, 0.53] for the BSI Anxiety scale. Note that although *Hypothesis 5* was supported for both CATs, Cohen's *d* suggested that the CAT Anxiety was somewhat less able than the BSI Anxiety scale to discriminate between patients with and without a primary anxiety diagnosis. For the Depression scales, however, we found the opposite: the CAT Depression was somewhat better able than the BSI Depression scale to distinguish between patients with and without a primary depression diagnosis.

In support of *Hypothesis 6*, Pearson's pretest to retest correlation coefficients differed less than 0.10 points between the CAT Depression (r = 0.54) and BSI Depression scales (r = 0.53). For the Anxiety scales, however, *Hypothesis 6* was rejected because the correlation coefficient for CAT (r = 0.40) was more than 0.10 points lower than that for the BSI (r = 0.56).

Overall, *Hypotheses 1-6* were supported for the CAT Depression. For the CAT Anxiety, *Hypotheses 1-5* were supported and *Hypothesis 6* was rejected. Consequently, construct validity was considered sufficient for both PROMIS CATs as more than 75% of the hypotheses were supported.

Table 6.1 Pearson's correlation coefficients between the PROMIS CATs and BSI subscales for the pretest and retest scores.

		Dep pre		Dep re		Anx pre		Anx re	
Scale	Instrument	CAT	BSI	CAT	BSI	CAT	BSI	CAT	BSI
Dep pre	CAT	1.00							
	BSI	**0.83**	1.00						
Dep re	CAT	0.54	0.46	1.00					
	BSI	0.51	0.53	**0.87**	1.00				
Anx pre	CAT	0.66	0.58	0.33	0.31	1.00			
	BSI	0.48	0.55	0.27	0.30	**0.76**	1.00		
Anx re	CAT	0.42	0.34	0.78	0.71	0.40	0.38	1.00	
	BSI	0.38	0.37	0.64	0.69	0.45	0.56	**0.81**	1.00
Som pre	BSI	**0.44**	0.48	0.32	0.32	**0.53**	0.63	0.33	0.45
Som re	BSI	0.35	0.35	**0.55**	0.59	0.37	0.41	**0.59**	0.70
Obs pre	BSI	**0.56**	0.66	0.39	0.42	**0.60**	0.62	0.37	0.46
Obs re	BSI	0.42	0.42	**0.64**	0.71	0.33	0.35	**0.67**	0.73
Hos pre	BSI	**0.39**	0.45	0.28	0.32	**0.38**	0.43	0.28	0.37
Hos re	BSI	0.28	0.27	**0.47**	0.53	0.21	0.25	**0.50**	0.56
Pho pre	BSI	**0.48**	0.53	0.37	0.38	**0.58**	0.66	0.41	0.50
Pho re	BSI	0.37	0.36	**0.60**	0.64	0.41	0.45	**0.66**	0.77
Par pre	BSI	**0.46**	0.52	0.28	0.32	**0.44**	0.47	0.28	0.37
Par re	BSI	0.34	0.37	**0.53**	0.59	0.27	0.29	**0.54**	0.60
Psy pre	BSI	**0.62**	0.73	0.37	0.43	**0.56**	0.56	0.32	0.39
Psy re	BSI	0.45	0.49	**0.72**	0.83	0.33	0.34	**0.68**	0.71
Int pre	BSI	**0.51**	0.62	0.34	0.41	**0.47**	0.52	0.31	0.42
Int re	BSI	0.36	0.38	**0.61**	0.69	0.30	0.32	**0.62**	0.70

Note. pre = pretest; re = retest; Dep = depression; Anx = anxiety; Som = somatization; Obs = obsessive-compulsive; Hos = hostility; Pho = phobic anxiety; Par = paranoid ideation; Psy = psychoticism; Int = interpersonal sensitivity; all correlations deviate statistically significantly from zero; correlations used to assess construct validity are presented bold-faced.

6.4.3 Responsiveness

Table 6.2 displays Pearson's correlation coefficients between the change scores of the PROMIS CATs and BSI subscales. In support of *Hypothesis 7*, the correlation coefficients between the CATs and matching BSI scales were above 0.50 for both Depression (r = .78) and Anxiety scales (r = .72). Furthermore, in support of *Hypothesis 8*, the correlation coefficients between the CATs and other BSI scales were at least 0.10 points below those between the CATs and matching BSI scales. Note that, similarly to the pretest and retest scores, the CATs correlated lower with the other BSI scales than the BSI Depression and Anxiety scales did.

Table 6.2 Pearson's correlation coefficients between the change scores of the PROMIS CATs and BSI subscales.

	Depression		Anxiety	
Scale	CAT	BSI	CAT	BSI
CAT Dep	1,00			
BSI Dep	**0,78**	1,00		
CAT Anx	0,67	0,61	1,00	
BSI Anx	0,55	0,64	**0,72**	1,00
BSI Som	**0,37**	0,49	**0,45**	0,59
BSI Obs	**0,46**	0,63	**0,60**	0,66
BSI Hos	**0,35**	0,46	**0,41**	0,44
BSI Pho	**0,44**	0,55	**0,48**	0,64
BSI Par	**0,46**	0,52	**0,47**	0,53
BSI Psy	**0,59**	0,71	**0,60**	0,62
BSI Int	**0,51**	0,64	**0,54**	0,60

Note. Dep = depression; Anx = anxiety; Som = somatization; Obs = obsessive-compulsive; Hos = hostility; Pho = phobic Anxiety; Par = paranoid Ideation; Psy = psychoticism; Int = interpersonal sensitivity; all correlations deviate statistically significantly from zero; correlations used to assess responsiveness are bold faced.

In support of *Hypothesis 9*, the *ES* value for the CAT Depression (pretest, $M = 64.5$, $SD = 6.8$; retest, $M = 60.8$, $SD = 8.1$; $ES = 0.55$, 95% *CI* [0.41 – 0.69]) was not more than 0.10 points lower than that for the BSI Depression scale (pretest, $M = 2.07$, $SD = 0.98$; retest, $M = 1.54$, $SD = 1.06$; $ES = 0.54$, 95% *CI* [0.40 – 0.68]). The *ES* value for the CAT Anxiety (pretest, $M = 67.2$, $SD = 6.4$; retest, $M = 63.1$, $SD = 7.5$; $ES = 0.64$, 95% *CI* [0.50 – 0.78]) was more than 0.10 points *higher* than that for the BSI Anxiety scale (pretest, $M = 1.79$, $SD = 0.93$; retest, $M = 1.32$, $SD = 0.92$; $ES = 0.50$, 95% *CI* [0.36 – 0.64]), which was also in support of *Hypothesis 9*.

Overall, *Hypotheses 7–9* were supported for both PROMIS CATs, indicating sufficient responsiveness. Under the assumption of measuring similar constructs, the CAT Anxiety even showed a higher responsiveness than the BSI Anxiety scale.

6.4.4 Utility of change indicators

Table 6.3 displays the percentages of (dis)agreement between the PROMIS CATs and matching BSI subscales for the four categories based on reliable change and *CSC*. In support of *Hypothesis 10*, the s^* statistic was lower than 0.60 for both CATs (Depression, $s^* = 0.53$, 95% *CI* [0.46 – 0.59]; Anxiety, $s^* = 0.50$, 95% *CI* [0.36 – 0.64]), the percentage of agreement was lower than 80% for both CATs (Depression, 11 + 6 + 54 + 1 = 72%; Anxiety, 11 + 3 + 52 + 1 = 67%), and less patients were categorized as unchanged by the CATs (Depression, 3 + 2 + 54 + 3 = 62%; Anxiety, 1 + 3 + 52 + 2 = 58%) relative to the BSI scales (Depression, 5 + 6 + 54 + 4 = 69%; Anxiety, 10 + 8 + 52 + 6 = 76%). These results suggest that, under the assumption of measuring similar constructs, change categorizations of the PROMIS CATs are substantially different from those of matching BSI scales, and the PROMIS CATs categorize more patients

as actually changed. Note that the difference between the CATs and BSI scales in the percentage of unchanged patients was larger for Anxiety than for Depression.

Table 6.3 Percentages of (dis)agreement between the PROMIS CATs and matching BSI subscales on the categories based on reliable change and CSC.

	CAT Depression				CAT Anxiety			
BSI	Recovered	Improved	Unchanged	Deteriorated	Recovered	Improved	Unchanged	Deteriorated
Recovered	11%	4%	3%	0%	11%	2%	1%	0%
Improved	1%	6%	2%	0%	2%	3%	3%	0%
Unchanged	5%	6%	54%	4%	10%	8%	52%	6%
Deteriorated	0%	0%	3%	1%	0%	0%	2%	1%

Note. The percentages add up to 101% for the Anxiety scales due to rounding.

6.5 Discussion

This was the first study in the Netherlands in which PROMIS CATs were administered. We evaluated construct validity, responsiveness, and utility of change indicators of the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CATs in a clinical sample, by comparing them with the Dutch BSI subscales. In line with other studies that used different legacy instruments (Kroenke et al., 2019; Pilkonis et al., 2014), we found that both PROMIS CATs showed sufficient construct validity, responsiveness, and utility of change indicators. More specifically, we found that the CATs measured similar constructs as matching BSI scales. Under the assumption of measuring similar constructs, the CAT Depression also showed a similar responsiveness relative to the BSI Depression scale. For the CAT Anxiety, we even found a higher responsiveness compared to the BSI Anxiety scale, which may suggest that the CAT Anxiety is more able to detect change. Finally, both CATs showed a substantial disagreement with matching BSI scales in change categorizations; the CATs categorized the change scores of more patients as changed, which may suggest that the CATs are more able to detect actual change. Based on these findings, the PROMIS CATs may be considered an improvement over matching BSI scales as tools for reviewing treatment progress with patients.

The findings of this study are based on the assumption that the BSI is an adequate comparator for the PROMIS CATs. It should be noted, however, that comparison instruments always differ to some extent, possibly due to differences in the methods used for test construction. It has been shown that different methods may amount to very different compilations of aspects on which a test performs well (Oosterveld et al., 2019). Identifying differences between the PROMIS CATs and the BSI may therefore help to explain some of the results in this study. First, the instruments differ in their underlying measurement theory and administration method. The PROMIS CATs were developed under an IRT model (Embretson & Reise, 2000), and use item banks to select and administer items that can differ between respondents and measurement occasions. The BSI was developed under the CTT model (Lord

& Novick, 1968), and uses a fixed number of items for all respondents and measurement occasions. Second, the PROMIS CATs provide a measurement error estimate for each individual test taker while the BSI scales only provide a single estimate of the standard error of measurement for all test takers. Third, the PROMIS CATs use response categories based on frequency (*never* to *always*) while the BSI uses response categories based on severity (*not at all* to *extremely*). Fourth, the PROMIS CATs use norm-based interval T-scores based on the US general population and the EAP estimator (Cella et al., 2010) while the BSI uses ordinal Dutch raw average scores. This means that the PROMIS CATs use prior information (i.e., the standard normal distribution) and the reliability of the test to improve the estimated score, whereas the BSI uses average scores without consideration of prior information (Bock, 1997). Finally, the PROMIS CATs were primarily developed for universal application in different populations, whereas the BSI was primarily developed for clinical populations. In case of the BSI Anxiety scale, the main focus was even more specific: patients with high anxiety levels (Derogatis et al., 1973).

In the next paragraphs, we provide possible explanations for the results based on the differences between the PROMIS CATs and the BSI, and the design of this study. We start with the results that stood out most regarding our hypotheses: the lower pretest to retest stability for the CAT Anxiety which led to the rejection of *Hypothesis 6*, and the higher responsiveness of the CAT Anxiety while we expected a similar responsiveness (Kroenke et al., 2019). Actually, both findings are in fact associated because stability is the opposite of change. To clarify this, consider that a scale's degree of stability is related to the variation in change scores of that scale: perfect stability results from all change scores being equal while low stability results from a large variation in change scores. The degree of variation in change scores may in turn be related to the degree of responsiveness. After all, higher responsiveness enables more space to be used on the scale, which can result in a larger variation of change scores. We therefore suspected that the CAT Anxiety showed a larger variation in change scores than the BSI Anxiety scale, which was confirmed by an additional analysis using *Z*-scores for both scales (not shown herein): the *SD* of the change scores was 1.20 for the CAT Anxiety and 0.94 for the BSI Anxiety scale. Consequently, it may be that the lower stability of the CAT Anxiety was to be expected, assuming a higher responsiveness.

We found two possible explanations for the unexpected finding of the CAT Anxiety having a higher responsiveness than the BSI Anxiety scale. First, the choice of item parameters influenced the results. We concluded this by recalculating the T-scores with DF item parameters (Flens et al., 2017, 2019) and re-evaluating responsiveness. The results (not shown herein) indicated that the pre-post effect size for the CAT Anxiety was somewhat smaller for DF item parameters (*ES* = 0.58) compared to US item parameters (*ES* = 0.64). Thus, would we have used the DF item parameters to calculate the T-scores, we would not have concluded that the CAT Anxiety was more responsive than the BSI Anxiety scale, but instead that they were similarly responsive. This difference was to some extent a consequence of the numerator in the *ES* formula (i.e., the *M* pretest T-score minus the *M* retest T-score; DF parameters = 4.00; US parameters = 4.06), but especially of the denominator (i.e., the *SD* of the pretest T-scores; DF parameters = 6.85; US parameters = 6.39). Apparently, DF item parameters yield a somewhat more conservative estimation of *ES* due to the larger range in pretest scores. This finding is

relevant for the discussion regarding the choice of appropriate item parameters (i.e., US parameters, country-specific parameters, or international parameters; Elsman et al., 2022; Terwee et al., 2021; van Bebber et al., 2018).

Second, the degree of longitudinal measurement invariance (LMI) may have influenced the degree of responsiveness. A set of items is said to show sufficient LMI when it measures one or more constructs in the same way over time. This means that changes in respondents' scores over time can entirely be attributed to changes *within* the construct(s) measured by the set of items (Fried et al., 2016; Liu et al., 2017). A previous study using full item bank data of Dutch patients with mood and anxiety disorders showed that the degree of LMI was sufficient in both PROMIS Depression and Anxiety item banks, but also that it was somewhat smaller in the PROMIS Anxiety item bank (Flens et al., 2021). Similarly, the degree of LMI may differ between the PROMIS CATs and matching BSI scales, which may have affected the degree of responsiveness (and perhaps other results as well). To investigate this, the BSI should be studied on LMI too, which was not within the scope of this study.

In addition, there were some findings of smaller importance in this study. First, Pearson's correlation coefficients showed that the PROMIS CATs had a lower association with the other BSI scales than the BSI Depression and Anxiety scales did. We found this for the pretest (*Hypothesis 3*), retest (*Hypothesis 4*) and change scores (*Hypothesis 8*), which may be somewhat expected when considering that the BSI scales have more in common with each other than with the CATs. Additionally, the (partially fixed) order in which the instruments were administered may have led to differences in respondent behavior (e.g., due to measurement fatigue, context effects, or order effects; Windle, 1954). As the BSI was always administered last, this may even have influenced other results as well. Unfortunately, the questionnaire-software of Dimence Group did not allow for further alternation between the instruments. For future studies, it is suggested that both PROMIS CATs and legacy instruments are alternated.

Second, Cohen's *d* showed that the CAT Anxiety pretest scale was somewhat less able than the BSI Anxiety pretest scale to discriminate between patients with and without a primary anxiety diagnosis (*Hypothesis 5*). For the Depression scales, however, we found the opposite for patients with and without a primary depression diagnosis. These findings may be explained by the item content of the scales. The PROMIS Anxiety item bank includes items that may be more appropriate for specific anxiety diagnoses such as an obsessive-compulsive disorder, phobia, or social anxiety. Consequently, the CAT Anxiety may select items that are less relevant for patients with other anxiety diagnoses, possibly leading to a somewhat lower latent trait level. The BSI Anxiety scale, on the other hand, includes mostly general anxiety symptoms. In this case, scores may be somewhat less affected because the administered items are relevant for most anxiety diagnoses. As a result, the CAT Anxiety's ability to discriminate between patients with and without a primary anxiety diagnosis may be somewhat lower than that of the BSI Anxiety scale. In contrast, this explanation may not apply to the CAT Depression as mood disorders may be less diverse in their manifestation than anxiety disorders. In this case, the selection of items from a larger item bank may lead, relative to administering a small fixed item set, to a somewhat better discrimination between patients with and without a primary depression diagnosis.

Third, there may be some method effects in the assessment of utility of change indicators (*Hypothesis 10*). For example, we used test-retest reliability instead of Cronbach's α for calculating the *RCI* for the BSI to account for variance in scores over time. Fortunately, an additional analysis (not shown herein) showed that our conclusions remained the same when using Cronbach's α based on the pretest of this study. In addition, the cutoff for *CSC* was calculated as the point halfway the general and clinical population (taking into account the variance in scores as well). A possible limitation of this method is that we had to use general population statistics from different samples for the PROMIS CATs and the BSI. Consequently, the results may have been affected by the degree of representativeness of these samples. For example, the general population samples were collected with stratified sampling for both CATs and BSI, but the sample used for the CATs accounted for more demographics variables than the sample used for the BSI (i.e., gender, age group, education, ethnicity, and region vs. gender and size of the city of residence), had a larger sample size (N = 1,002 vs. N = 200) and was collected more recently (2016 vs. 2005). Based on these differences, we could have chosen another cutoff for *CSC* that is calculated using data of the current study only. In this method, *CSC* is defined as a patient moving more than 2 *SD's* from the mean of the clinical sample (Jacobson et al., 1984). Fortunately, an additional analysis (not shown herein) showed once more that our conclusions remained the same. These findings indicate that method effects were not meaningful for the assessment of utility of change indicators.

Last, the PROMIS CATs used a stopping rule that combined measurement precision and an upper limit of administered items while the BSI Depression and Anxiety scales always administered six fixed items. Consequently, we could not eliminate any concern that findings are due to different test lengths. This could have been solved by using a stopping rule that always administered six items according to the CAT algorithm, but we preferred to use a stopping rule that most likely will be used in clinical practice to provide test users with practical information to choose instruments. As a result, they can make their own trade-off between efficient measurement and reliable measurement of the PROMIS CATs and the BSI, based on the information available.

In this study, both PROMIS CATs were shown to be sufficiently efficient, valid, and responsive relative to the BSI subscales. For utility of change indicators, we found modest improvements for the PROMIS CATs compared to matching BSI scales, which is likely due to the PROMIS methodology. Both PROMIS CATs use state of the art CAT administration, resulting in a highly relevant selection of items that is tailored to each respondent's severity level. Furthermore, CAT ensures that each administration meets the minimally required measurement precision, by which the number of administered items is allowed to vary among respondents. Consequently, measurement is both efficient and reliable for a large range of severity levels (Flens et al., 2017, 2019). The BSI subscales, however, use fixed item sets with a small number of items. As a result, measurement precision can vary among respondents (Reise & Waller, 2009) and may be generally lower than that of the PROMIS CATs (Pilkonis et al., 2014). In addition, PROMIS CATs provide a measurement error estimate for each individual test taker while the BSI subscales only provide a single estimate of the standard error of measurement for all test takers. Consequently, change indicators may be more accurate for the PROMIS CATs compared to the BSI (Brouwer et al., 2013; Mancheño et al., 2018). Based on

this, the PROMIS CATs may be considered an improvement over matching BSI scales as tools for reviewing treatment progress with patients.

For current BSI users, other results may also need to be considered to decide whether to change instruments. First, the responsiveness of the CAT Anxiety was somewhat higher than that of the BSI Anxiety scale, which was unexpected considering the results of previous studies (Kroenke et al., 2019; Pilkonis et al., 2014). Second, the administration efficiency of the instruments was quite similar. The CAT Anxiety even administered somewhat more items on average (i.e., 8 items) relative to the BSI Anxiety scale (i.e., 6 items). Note, however, that relative to the CAT Depression, the CAT Anxiety also categorized a larger degree of patients as changed compared to the matching BSI subscale, which may be due to the extra items. Finally, our study design may have disadvantaged one of the study measures by always administering the BSI last, increasing the uncertainty of the results. Based on these findings, it may not yet be appealing to all BSI users to make the transition to PROMIS CATs, especially considering that test users need to get used to new instruments, which may be experienced as a burden.

When BSI users are sufficiently convinced to change instruments, the PROsetta Stone® initiative offers the possibility to convert BSI Depression scores into PROMIS CAT Depression scores for an easier transition (www.prosettastone.org/new-page-1-1; Kaat et al., 2017). Using PROMIS instruments also has additional advantages for practice that are beyond the scope of this study. For example, PROMIS instruments are universally applicable in a wide range of populations whereas the BSI is mostly used in populations that primarily suffer from mental health problems (Beleckas et al., 2018; Lizzo et al., 2019; Papuga et al., 2018; Scholle et al., 2018; Wagner et al., 2015). PROMIS scores may even be compared across countries to learn from each other's practice (Elsman et al., 2022; Terwee et al., 2021; van Bebber et al., 2018). In addition, test users have access to numerous other PROMIS (CAT) instruments measuring different constructs of a large part of the health spectrum (for more details, see www.healthmeasures.net/explore-measurement-systems/promis/obtain-administer-measures). This means that PROMIS users have more flexibility in administering a set of instruments that specifically fits the patient's treatment goals, instead of being bound to BSI subscales that may not all have to be relevant for a patient.

Strengths of this study are the sample properties and the assessment procedure. The sample included only patients that completed the PROMIS CATs and the BSI on the same day for both pretest and retest, resulting in $N = 400$ while typically $N = 200$ is used for the performed analyses (e.g., Pilkonis et al., 2a014; Schalet et al., 2016). Furthermore, the response rate was substantial (i.e., 72.9%), and the composition of the sample (regarding gender, age, and pretest severity level) was representative for the mental health provider that collected the data. In contrast, the sample may lack representativeness for the Dutch clinical population because the data were not collected using stratified sampling. For example, the Dimence Group has many departments, covering urban and rural areas, albeit only in the east of the Netherlands. Consequently, few patients from other regions in the Netherlands were included, possibly affecting the representativeness of the sample (Dieperink et al., 2008). In addition, the patients of this study showed somewhat more severe symptoms at the start of treatment than the patients

used for calibrating the PROMIS item banks for Depression and Anxiety (Flens et al., 2017, 2019), possibly affecting the representativeness of the sample too.

We have several suggestions for future research. The tentative rules of thumb that were used for some of the analyses need to be evaluated in a (simulation) study to assess whether they correspond sufficiently to the suggested interpretations. Also, our sample consisted mostly of patients with a depression or anxiety disorder (i.e., 85%). Because the PROMIS CATs for Depression and Anxiety may also be relevant for patients with other conditions, such as diabetes (Lloyd et al., 2000), cancer (Singer et al., 2010), cardiovascular diseases (Hare et al., 2014), and other mental health disorders (e.g., attention deficit disorder, somatoform disorder, personality disorder; Clarke & Kissane, 2002; Frank, 1974), it is suggested for future studies to re-evaluate the investigated psychometric properties for these conditions as well.

In addition, it is suggested to compare the DF PROMIS CATs to other legacy instruments, such as the PHQ-9 (Kroenke et al., 2001), the GAD-7 (Spitzer et al., 2006), and the Mood and Anxiety Symptom Questionnaire (MASQ; Flens et al., 2016; Watson & Clark, 1991). In a previous study, the US CAT Depression was compared to the PHQ-9 and the CESD (Pilkonis et al., 2014). Similar to our study, construct validity was found to be sufficient relative to the legacy instruments. One unexpected finding, however, was that the CAT Depression displayed the smallest pretest to retest effect size. The authors suggested that this was likely a consequence of the decreased variance in the legacy instruments due to floor effects. Furthermore, they argued that such a result raises the possibility that commonly used instruments may overestimate effect sizes. Fortunately, floor effects for the BSI scales were of minor importance in this study. In an additional analysis (not shown herein), we found for both BSI Depression and Anxiety scales that approximately 5% of the patients had a retest score of 0. However, floor effects may be generally larger when all retests are administered at the end of treatment, possibly affecting the responsiveness and the utility of change indicators.

Following this line of reasoning, the wide range in the pretest to retest interval may also have affected the results of this study. It may be, for example, that the results will be different for respondents with a small pretest to retest interval compared to respondents with a high pretest to retest interval (e.g., due to differences in floor effects in the BSI scales). To investigate this, we split the study sample into two equal halves based on the median pretest to retest interval, and repeated the analyses of this study (not shown herein). We found that our conclusions remained the same in both subsamples, indicating that the length of the pretest to retest interval did not have a substantial effect. However, it may be recommended for follow-up research to additionally evaluate this for patients that are reassessed over even longer time-intervals. Note, for example, that the retest scores in this study were still somewhat high, and the change scores somewhat low. Therefore, the question remains whether the results will also be similar when the change scores are larger.

In this study, we compared the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CAT with the nine subscales of the BSI in a clinical sample. Overall, our study suggests that the PROMIS CATs measure the same constructs as matching BSI scales, were at least as responsive, and categorized the change scores of more patients as actually

changed. Based on these findings, the PROMIS CATs may be considered a modest improvement over matching BSI scales as tools for reviewing treatment progress with patients.