Universiteit Leiden
The Netherlands

Psychopathology and health-related quality of life as patient-reported treatment outcomes: evaluation of concordance between the Brief Symptom Inventory (BSI) and the Short Form-36 (SF-36) in psychiatric outpatients.
Beurs, E. de; Carlier, I.; Hemert, A. van

Psychopathology and health-related quality of life as patient-reported treatment outcomes: evaluation of concordance between the Brief Symptom Inventory (BSI) and the Short Form-36 (SF-36) in psychiatric outpatients

Edwin de Beurs[1,2] · Ingrid Carlier[3] · Albert van Hemert[3]

Abstract

Purpose Treatment outcome for common psychiatric disorders, such as mood and anxiety disorders, is usually assessed by self-report measures regarding psychopathology [e.g., via Brief Symptom Inventory (BSI)]. However, health-related quality of life [as measured by the 36-item Short-Form Health Survey (SF-36)] may be a useful supplementary outcome domain for routine outcome monitoring (ROM). To date, the assessment of both outcomes has become fairly commonplace with severe mental illness, but this is not yet the case for common psychiatric disorders. The present study examined among outpatients with common psychiatric disorders whether aggregate assessments of change across treatment regarding psychopathology and health-related quality of life yield similar results and effect sizes.

Methods We compared treatment outcome on the BSI and the SF-36 in a sample of 13,423 outpatients. The concordance of both instruments was assessed at various time points during treatment.

Results Scores on both instruments were associated, but not so strongly to suggest they measure the same underlying construct. The SF-36 scales presented a varied picture of treatment outcome: understandably, patients changed more on the mental component scales than on physical component scales. Outcome according to the BSI was quite similar to outcome according to scales of the SF-36 that showed the largest change.

Conclusions Although (mental health) scores on both instruments are associated, adding the SF-36 in addition to the BSI in treatment evaluation research produces valuable information as the SF-36 measures a broader concept and contains physical/functional component scales, resulting in a more complete clinical picture of individual patients.

Keywords BSI · SF-36 · Patient-reported outcome measures (PRO) · Routine outcome monitoring (ROM) · Treatment evaluation · Common mental disorders

Introduction

Measurement-based care (MBC) uses patient-reported rating scales in conjunction with evidence-based clinical practice guidelines to provide an objective assessment of patient progress over time to guide a more precise plan of care [1], prevent treatment failure [2], and collect data for quality management [3]. Through the use of patient-reported outcome, PROs [4], the patient's voice is being heard, quantified, and compared to normative data in a large variety of domains [5]. This is in line with international trends, where more emphasis is put on the value of health care in terms of outcome [6] and patients are granted a more prominent role [7].

Current health definitions involve at least three domains: physical, mental, and social health that should be prioritized in delivering health care [8, 9]. Ideally, health outcomes should include all three domains of health in a full cycle of care [10]. In somatic care (e.g., oncology), the use of MBC has become largely routine, and usually

✉ Edwin de Beurs
    edwin.de.beurs@arkin.nl

1   Department of Clinical Psychology, Leiden University,
    Leiden, Netherlands

2   Arkin Mental Health Institute, Amsterdam, Netherlands

3   Department of Psychiatry, Leiden University Medical Centre,
    Leiden, Netherlands

several health domains are measured (e.g., physical symptoms, functioning, anxiety, depression) [1]. In psychiatric care, on the other hand, MBC is less standard practice, due to several barriers, such as lack of agreement on key outcome domains and lack of empirical data on outcome measures [3]. Measurement often mainly focuses on mental health without including other measurement domains, such as functioning or wellbeing [11, 12].

In mental health care research, outcome is commonly assessed by comparing the severity of psychopathology before and after treatment with generic or disorder-specific instruments. In this context, a widely used instrument is the Brief Symptom Inventory, BSI [13, 14], which provides with its total score (the Global Severity index) information regarding the severity of general psychopathology as well as on specific symptoms, such as depression and anxiety. In assessing the severity of psychopathology, these instruments measure signs and symptoms of the disorder (e.g., those listed in the prevailing taxonomy of mental disorders, the DSM-5).

However, operationalization of outcome in mental health care by signs and symptoms has been criticized as too narrow and too much focused on deficits [10, 15]. Health is more than the absence of signs and symptoms. Patients' view on their health-related quality of life (HRQOL) offers a broader conceptualization and may yield a useful additional indicator of treatment outcome. HRQOL is defined as the quality of life relative to one's health or disease status, and it is commonly conceived as dynamic, subjective, and multidimensional [16]. This shift in emphasis is also reflected in the emergence of the recovery movement in psychiatry, with its distinction between clinical and personal recovery [17], and positive psychology [18, 19] as well as positive psychiatry movements [20].

An instrument for the assessment of HRQOL is the Short Form-36, SF-36 [21, 22], widely used in health care and mental health care. The existing literature on HRQOL in mental health care research is predominantly concerned with severe mental disorders, such as psychotic disorders [23–25], for which the measurement of HRQOL is seen as a necessary addition to other outcome domains such as psychopathology. For common mental disorders, such as mood- and anxiety disorders, the value of adding HRQOL to the assessment of treatment outcome is less well investigated. Assessment of HRQOL in mood disorders has been recommended [26], but little research comparing measures head-to-head has been done [27]. In a meta-analyses for anxiety disorders, Olatunji et al. [28] reported that Post-Traumatic Stress Disorder (PTSD) in particular was associated with decreased HRQOL. This finding was confirmed in a recent study with PTSD patients [29], demonstrating a strong association between the change in depression symptoms and

change in HRQOL, which could be expected as depression symptoms are incorporated in HRQOL measures.

Thus, there is a strong plea for a broader assessment of the benefits of mental health care than mere symptom relief [30]. Adding an outcome domain to signs and symptoms is especially valuable when a decrease in signs and symptoms correlates only moderately with increased health-related quality of life. This may be the case when change over time on both constructs does not occur in synchrony. Many patients fist show improvements in symptomatology, to be followed later on by an increase in health-related quality of life. However, the precise association between symptomatology and quality of life in MHC is still poorly understood [31].

The present study used routine outcome monitoring (ROM) data of outpatients with common mental disorders to investigate and compare both outcome domains. Longitudinal data of the BSI and the SF-36 were compared, their correlation assessed, and—more importantly—the concordance between a decrease in score on the BSI with an increase in score on the SF-36 over time was established. We investigated whether the overall magnitude and the pace of change over time was similar in both domains. After all, a common hypothesis is that therapeutic change is first manifest on symptoms (in this case measured by the BSI), to be followed later by improved functioning or HRQOL [32, 33]. We analyzed this issue using a subset of the sample with four repeated assessment per patient, which enabled us to test this hypothesis regarding the dyssynchrony of response on the two outcome domains of psychopathology and HRQOL.

In sum, the main objective of the study was to investigate and compare the responsiveness of two outcome measures: a symptom checklist (BSI) and a HRQOL measure (SF-36). An asynchronous response pattern on both constructs was hypothesized: Symptoms decrease first, followed later on by an increase in quality of life.

Methods

Procedure

The collection of data is described in detail by de Beurs et al. [34]. Here, we provide a brief description. For this study, we used data from the Dutch Center of Routine Outcome Monitoring (COROM) from 2003 to 2013. ROM data have been collected at the Leiden University Medical Centre (LUMC, Department of Psychiatry) and at the Mental Health Care Provider GGZ Rivierduinen (the main service provider in the northern half of the South-Holland province, an area with 1.1 million inhabitants). Outpatients with mood-, anxiety-, and/or other psychological problems were referred for treatment by their general practitioner. All

patients were informed that ROM was part of the intake process. They were interviewed by a research nurse about their psychiatric symptoms with a semi-structured diagnostic interview (MINI-Plus [35, 36]). Next, patients completed a set of generic and disorder-specific questionnaires. The selection of disorder-specific questionnaires was based on the outcome of the MINI-Plus [35]. During treatment, they were assessed every three to four months with a comprehensive battery of self-report instruments and rating scales. Per sampling round, the number of patients decreased by 50%, partly due to the completion of treatment, partly because of no-show (after repeatedly being contacted by mail, email, and phone) of the patient at the test session. About half the patients discontinued ROM, but continued treatment. For the other half, the last available assessments coincided with ending the treatment and can be considered as the post-test assessment [34]. The mean measurement interval after the pretest was a little over 6 months, from the 2nd to 3rd and from the 3rd to the 4th measurement a little short of 6 months (see Table 2).

Study population

Initially, data of 13,811 psychiatric outpatients were available; 5826 were retested at least once with the BSI and the SF-36 and we used this sample for analyses regarding responsiveness (628 cases were excluded, based on negative retest intervals or a retest interval < 2 weeks as these were considered administrative errors; also, cases where the assessment on the BSI and the SF-36 were more than a week apart were excluded). Table 1 presents demographic and clinical data on the sample with pre- and retest data and the sample with four assessments. Most patient suffered from a Major Depressive Disorder or an anxiety disorder (singular or comorbid). A substantial number of patients (about 20 to 25%) did not meet criteria for a current DSM-IV disorder according to the MINI-Plus. Table 2 presents data regarding loss to follow-up; for 1463 patients at least four assessments were available. The pretest scores of the sample with a single retest and the subsample with four assessments were similar.

Patients were treated according to evidence-based guidelines with pharmacological and/or psychological treatments (mainly Cognitive Behavior Therapy [37]). The Medical Ethical Committee of the LUMC approved the general study protocol regarding ROM, in which ROM is considered integral to the treatment process (no written informed consent is institutionally required for the analysis of anonymized data). A comprehensive protocol (Psychiatric Academic Registration Leiden database) was used, which safeguarded the anonymity of participants and ensured proper handling of the data. All participants gave permission for the anonymized use of their data for scientific study.

Instruments

Mini-International Neuropsychiatric Interview Plus

The MINI-Plus [35] is a fully structured diagnostic interview that assesses DSM-IV criteria for the main psychiatric disorders (current/lifetime) such as mood-, anxiety-, somatoform-, substance use-, psychotic-, eating-, conduct, attention-deficit/hyperactivity, adjustment-, and anti-social personality disorders. It is organized in 26 modules: affirmative answers to screening questions are explored by establishing the presence of additional diagnostic criteria. Excellent inter-rater and test–retest reliabilities of the MINI have been established [35].

Short Form-36

The SF-36V1 [21, 22], also known as the RAND-36 or MOS-36, is a widely used instrument for measuring HRQOL or functional health status. The 36 items refer to the last 4 weeks, except for the scale Physical Functioning (PF) ("at this moment") and General Health Perceptions (GHP) ("in general"). The items have measurement scales of different lengths (2-, 3-, 5-, and 6-point scales), which are converted to a 0–100 scale before they are averaged into scale scores. Items are allocated to eight subscales: *Physical Functioning (PF), Role limitations due to Physical problems (RP), Bodily Pain (BP), General Health perceptions (GH), Vitality (VT), Social functioning (SF), Role limitations due to Emotional problems (RE), and general Mental Health (MH)*. In addition, scores for two orthogonal health components can be calculated from the SF-36 scales: the physical and the mental component, respectively, PCS (primarily reflecting PF, RP, BP, and GH) and MCS (primarily reflecting VT, SF, RE, and MH). The component scores are based on Principal Component Analysis of SF-36 scale scores derived from the American general population as recommended by Ware et al. [38]. Together, they explain 80 to 85% of the variance in scale scores and the components have proved reliable and valid. Component scores are created by multiplying standardized scale scores, based on Dutch norms [22], with the component score coefficients of the two-component solution [39]. A higher score means better health.

Brief Symptom Inventory

The BSI [13, 14] is one of the most frequently used general symptom measures in mental health care regarding psychopathology. It consists of 53 items (a selection of the best performing items of the SCL-90, the precursor of the BSI), each describing a "problem" (complaint or symptom). The respondent is asked to indicate "how he/she has been affected by this problem, the past week including

Table 1 Demographic and pretest clinical characteristics of the sample with at least two assessments ($n = 5826$) and the sample with four assessments ($n = 1463$) and posttest scores for the latter sample

Diagnosis (DSM-IV):	Twice assessed		Four times assessed			
	Pretest		Pretest		Posttest	
	N	%	N	%		
Gender (female)	3674	63.1	929	63.5		
Age (M, SD)	39.1	12.7	39.1	13.0		
Diagnosis						
Singular mood disorder	1717	29.5	467	31.9		
Singular anxiety disorder	1295	22.9	297	20.3		
Comorbid mood/anxiety	1164	20.0	353	24.1		
Singular other disorder[a]	43	0.7	13	0.9		
Comorbid mood/other[a]	55	0.9	18	1.2		
Comorbid anxiety/other[a]	30	0.5	9	0.6		
Comorbid mood/anxiety/other[a]	64	1.1	20	1.3		
No DSM-IV disorder	1457	25.0	288	19.7		
	M	SD	M	SD	M	M
BSI: TOT	1.23	0.70	1.32	0.69	0.83	0.65
DEP	1.56	1.00	1.69	0.99	0.96	0.89
ANX	1.39	0.94	1.49	0.95	0.90	0.81
SF-36: PCS	49.1	10.7	48.6	10.7	49.3	9.9
MCS	28.0	11.3	26.5	10.9	36.8	12.5
PF	76.4	23.4	74.7	23.2	79.8	22.2
RP	40.3	41.0	37.4	40.3	53.0	41.9
BP	68.3	27.1	67.4	26.8	72.6	25.1
GH	53.1	21.0	51.2	20.2	57.9	21.3
VT	36.1	18.4	33.6	17.7	47.4	20.7
SF	47.4	26.6	44.3	26.0	63.7	26.6
RE	30.7	37.5	27.3	36.0	50.9	42.2
MH	44.1	18.8	41.5	18.3	57.7	19.9

BSI: TOT total score, *DEP* Depression, *ANX* Anxiety, *SF-36: PCS* Physical Component Score, *MCS* Mental Component Score, *PF* Physical Functioning, *RP* Role limitations Physical, *BP* Bodily Pain, *GH* General Health perceptions *VT* Vitality, *SF* Social Functioning, *RE* Role limitations Emotional, *MH* Mental Health

[a]Somatoform disorder, eating disorder, addictive disorder, adult attention-deficit/hyperactivity disorder

Table 2 Number of patients at each assessment and the length of various time intervals

	1st	2nd	3rd	4th assessment
Entire dataset	13,811	6579	3392	1723
Both BSI and SF-36 present	13,432	6454	3329	1694
With outliers removed[a]	12,806	5826	2946	1463
Reassessment after M (SD) days		197 (142)	172 (117)	170 (110)

[a]628 Cases (9.7%) were removed, due to negative, extremely short or extremely long time intervals

today" on a 5-point Likert scale, ranging from $0 =$ none to $4 =$ very much. The BSI includes nine scales, each with four to six items: somatic symptoms, cognitive symptoms, interpersonal sensitivity, depression, anxiety, hostility, phobic symptoms, paranoia, and psychoticism. Furthermore, a total score can be calculated representing severity of general psychopathology. For the present study, the total score is used (*BSI–TOT*) as well as the Depression (*BSI-DEP*) and Anxiety (*BSI-ANX*) subscales. A higher score means more psychopathology. The BSI and the SF-36 are both patient-reported outcome (PRO) measures.

Statistical analyses

To get a similar direction in scores on the BSI and the SF-36, the SF-36 scales were reversed. In addition, scores on the BSI and the SF-36 scales were transformed to the same metric by subtracting the pretest mean score and dividing it by the pretest standard deviation (standard or Z-scores). Thus, for the entire sample, the initial score is $M = 0$ (SD = 1), and difference scores from pretest to subsequent assessments are of the same size as the within-group effect size (ES) estimator: $ES = (M_{pretest} - M_{posttest})/SD_{pretest}$) [40]. For ES, 0.20 indicates a small effect, 0.50 a medium effect, and 0.80 a large effect [41]. For the analysis of the longitudinal data, a subgroup with at least four assessments was selected, as this yielded the most optimal balance between the number of retests and loss to follow-up. Furthermore, we established the ES of the first assessment interval (ES_{1-2}) and for the interval between the first and the last available complete assessment of each patient in the entire study group (ES_{MAX}).

To assess construct validity of the BSI and the SF-36, we determined the correlation between scores at pretest and at the final assessment; to compare responsiveness we determined correlations between change scores (pre-to-retest difference) from the first to the second assessment and from the first to the final assessment [42]. To detect dissimilarity and asynchrony of response on the BSI and the SF-36, we compared the course of scores on both instruments over time with multivariate analysis of variance for repeated measurements. In this analysis, a significant interaction effect between time and instrument indicates a difference between instruments in change over time. For the instrument effect, after the omnibus test a "simple" contrast was tested, which compared the BSI total score with each SF-36 component and scale score. The repeated contrast for time compares the first with the second assessment, the second with the third, and so on. With four time points, it can be determined whether there is a temporal difference in change according to the two outcome measures (e.g., first a change in psychopathological symptoms (BSI), followed by a change in health-related quality of life (SF-36) in the ensuing time interval). The large number of observations per analysis (generally $n > 1500$) provides abundant statistical power to find differences. Consequently, also small differences will be statistically significant. Therefore, it is more important to look at the proportion of variance accounted for by each statistically significant effect [43]. For the between-instrument effect, the time effect, and the interaction effect, we present partial η^2 as provided by SPSS. For η^2, 0.01 indicates a small effect; 0.06 a medium effect; 0.14 a large effect [41].

Results

Concordance of the BSI–TOT with SF-36 scales

A first step to detect dissimilarity and asynchrony of response on the BSI and the SF-36 was to investigate the correlation between both measures. Table 3 shows Pearson correlation coefficients for the association of the total score on the BSI and the SF-36 scale scores. The correlation coefficients of the BSI with the component and scale scores were generally between $r = .40$ and .70, indicating that the BSI and the SF-36 measure different but related concepts. In accordance with its purported measurement aim, the SF-36 general Mental Health scale (MH) correlates the strongest with the BSI total score; second highest is the Mental Component Score (MCS). This applies to both the baseline data and the last available assessment. Correlation between change scores for the first measurement interval and the maximum measurement interval are lower, but change scores are still substantially associated. Again, the highest correlations are found between the BSI and the MCS and MH scales. Furthermore, results appear to stabilize over time, as correlations between component and scale scores are systematically lower at the first assessment versus the last available assessment, and the fist interval and the last interval.

Table 3 Correlation (product moment correlation coefficients) between BSI total score and the SF-36 component and scale scores at the first and the last available (nth) assessment and between difference scores

BSI total score	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
1st Assessment (n = 13,432)	.25	.66	.35	.33	.39	.43	55	.57	.47	.75
nth Assessment (n = 5826)	.36	.76	.44	.51	.48	.57	.66	.69	.60	.82
Difference score 1st interval	.14	.64	.27	.30	.26	.39	.54	.52	.40	.70
Difference score nth interval	.16	.66	.30	.34	.28	.42	.57	.57	.44	.72

All correlations are two-sided significant at $p < .001$

PCS Physical Component Score, *MCS* Mental Component Score, *PF* Physical Functioning, *RP* Role limitations Physical, *BP* Bodily Pain, *GH* General Health perceptions, *VT* Vitality, *SF* Social Functioning, *RE* Role limitations Emotional, *MH* Mental Health

Gradient over time

The data of the group with four assessments ($n = 1463$; mean interval of first to fourth measurement $M_{1–4} = 471.9$ days, SD $= 184.1$) were analyzed with repeated measures analysis to compare the course over time of the BSI total score with the two SF-36 component scores for physical and mental health (PCS and MCS). This revealed main effects of time ($[F(3, 1460) = 307.46; p < .001, \eta^2 = .387)$, instrument $[F(2, 1460) = 84.38; p < .001, \eta^2 = .104]$, and a significant interaction effect for instrument-by-time $[F(6, 1457) = 74.86; p < .001, \eta^2 = .236])$. On each consecutive assessment, scores diminished on the BSI and increased on the SF-36, indicating that, on average, patients improved over time. Table 4 presents the values of partial η^2 for main effects and interaction effect of the time-by-instrument factorial model. The second column presents partial η^2 for the overall effect, later columns present partial η^2 for repeated contrasts. Contrasts comparing the consecutive assessments (repeated contrast) show that the largest change occurs in the first assessment interval. The instrument effect was predominantly due to the BSI–PCS contrast (less change over time for the PCS). The significant time-by-instrument interaction was predominantly due to the difference in change according to the BSI and the PCS in the first interval. The statistical significance of the effects indicates that the BSI differed from the PCS for all intervals, but it differed only in the first interval significantly from the MCS (with a larger change on the MCS). Analyses of standardized mean differences (partial η^2) corroborate this finding. Figure 1 clarifies these findings graphically as it shows the course of scores over time for patients with at least four assessments. The size of η^2 for the time contrast effect in Table 4 indicates that the lines in Fig. 1 deviate significantly from a horizontal course (the greatest decrease in score is attained in the starting phase); the statistical significance and size of η^2 for the interaction
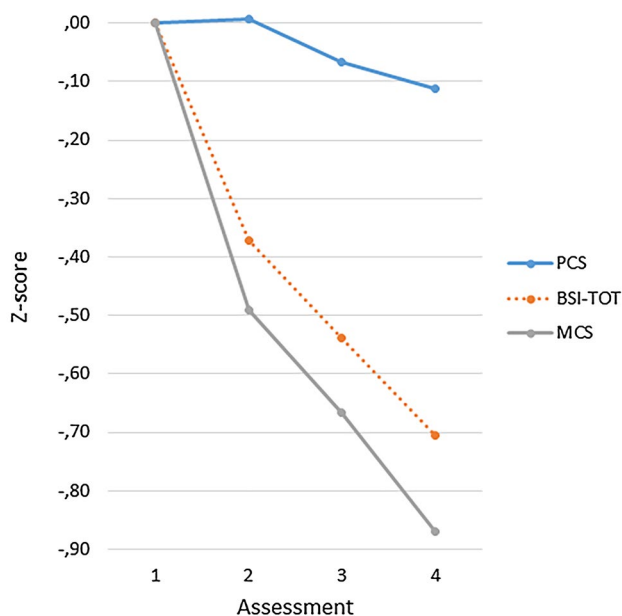

Fig. 1 Course over time of standardized *Z*-scores on the BSI total score and SF-36 Physical and Mental components for the group with four assessments ($n = 1453$)

effect of time * instrument indicates that the lines do not run parallel. The difference between the BSI and the PCS was substantial (partial $\eta^2 = .125$ for the first interval). In the first interval, there was a larger change on the MCS compared to the BSI, but no difference between both scores in the last two time intervals. Thus, our hypothesis of an asynchronous response between the BSI and the SF-36, with scores on the BSI decreasing first, is not supported by the findings.

A similar analysis of BSI total score and all subscales of the SF-36 revealed a main effect for time $[F(3, 1460) = 283.42; p < .001, \eta^2 = .368]$, a main instrument effect $[F(8, 1455) = 33.36; p < .001, \eta^2 = .155]$, and a

Table 4 Overview of effect sizes (partial η^2) for main and interaction effects of the analysis of variance for repeated measures using contrasts to compare BSI–TOT with the SF-36 component scores over four consecutive assessments (repeated contrast)

	Overall		Repeated contrast		
Time effect	**.224**		From 1–2	From 2–3	From 3–4
			.187	.055	.057
		Contrast			
Instrument effect	.084	BSI–TOT vs. SF-36–MCS	.033		
		BSI–TOT vs. SF-36–PCS	.078		
		Repeated contrast			
Time * instrument	.082	Contrast	From 1–2	From 2–3	From 3–4
		BSI–TOT vs. SF-36–MCS	.018	.000*	.002*
		BSI–TOT vs. SF-36–PCS	.125	.009	.019

BSI–TOT Brief Symptom Inventory Total Score, *SF-36–PCS* SF-36 Physical Component Score, *SF-36–MCS* SF-36 Mental Component Score

*Small effects and interaction effects were not statistically significant ($p > .10$); partial $\eta^2 = 0.01$ indicates a small effect, $\eta^2 = 0.06$ indicates a medium effect, and $\eta^2 = 0.14$ is a large effect (in bold typeface)

significant interaction effect for instrument-by-time [$F(24, 1439) = 26.59$; $p < .001$, $\eta^2 = .307$]. Pairwise comparison of the BSI total score with SF-36 scores reveals that the decrease in score on the BSI differs from all the SF-36 scales, except for the RE-scale (Role limitations due to Emotional problems). Figure 2 illustrates the pattern of change in BSI and SF-36 scale scores over time. It appears that two groups of SF-36 scales can be distinguished regarding how much change they detect. The scales primarily associated with the mental component (VT, SF, RE, and MH) show more change than the scales primarily associated with the physical component (PF, RP, BP, and GH). Figure 2 also shows change in depression and anxiety according to the BSI depression and BSI anxiety subscales, which is comparable to the BSI total score. Changes on the BSI scales are similar to the mental component scales. Furthermore, we inspected the scores of the samples with three ($n = 2946$) or two assessments ($n = 5786$) and found profiles of scores very similar to Figs. 1 and 2 for these larger samples. All in all, according to these analyses the hypothesized dyssynchrony is absent, the instrument-by-time interaction is merely due to less change on the scales measuring physical health.

Finally, Fig. 3 presents effect sizes (ES) for the difference in BSI total score and depression and anxiety subscales, the SF-36 scales, and component scores over the first measurement interval (ES_{1-2}) and the maximum interval from baseline to final assessment of the entire group (ES_{MAX}). Again, the scales BP, PF, GH, and RP show the least change
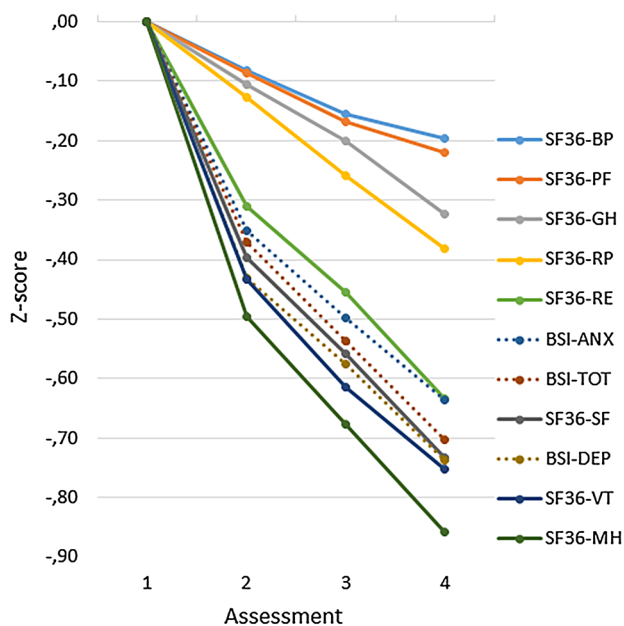

Fig. 3 Effect size (ES) of change in the first assessment interval (ES_{1-2}) and the maximum change (ES_{max}) according to the BSI–TOT (total score), two BSI subscales (DEP and ANX), the SF-36 Physical (PCS) and Mental Component Scores (MCS), and the eight SF-36 scales; the order in the legend corresponds to the ranking at the last assessment

($ES_{MAX} = 0.18$ to 0.38) and the scales MH, RE, SF, and VT show the most change ($ES_{MAX} = 0.64$ to 0.84); the change on the BSI total score ($ES_{MAX} = 0.62$) is closer to the latter group of scales than to the former. Change on the BSI is similar to scales of the SF-36 that demonstrate the largest change and the BSI depression and anxiety subscales show a somewhat larger change than the BSI–TOT.

Discussion

The main findings of this study are as follows. There is a high correlation between the BSI total score and component and scale scores of the SF-36, especially for the MH scale. This is in spite of the fact that the BSI and the SF-36 use different temporal instructions for the rated time frame: 1 week vs. 4 weeks, respectively. The substantial concordance between the BSI and the MH scale of the SF-36 is not surprising, given the content of the relevant SF-36 items ("Did you feel nervous?," "Did you feel downhearted and blue?"), which are very similar to BSI items ("Nervousness or shakiness inside," "Feeling blue"). The MH scale of the

Fig. 2 Course over time of standardized Z-scores on the BSI–TOT (total score), two BSI subscale scores (DEP and ANX) and eight SF-36 scales for the group with four assessments ($n = 1453$); the order in the legend corresponds to the ranking at the last assessment
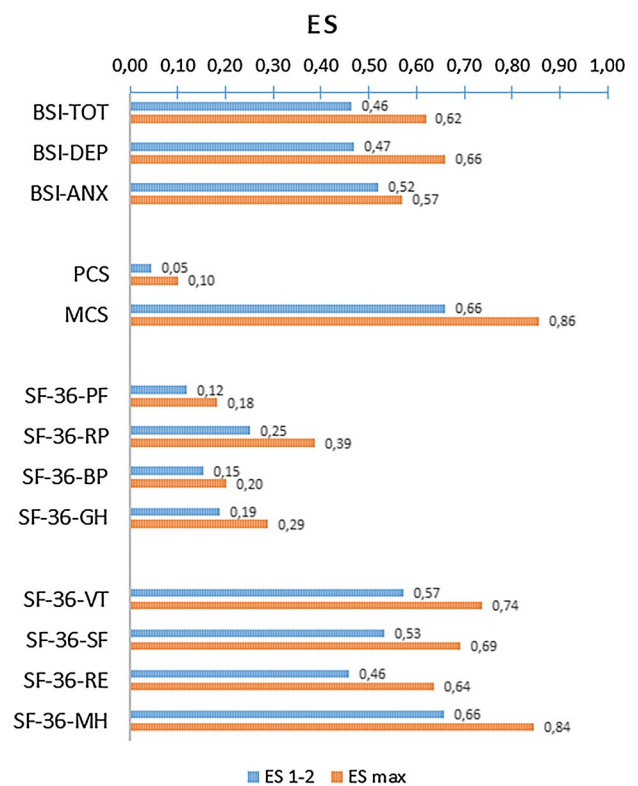
SF-36 demonstrated somewhat larger changes than the BSI total scale. A likely explanation for this finding is that the BSI total scale contains a substantial number of items with low relevancy for patients with mood or anxiety disorders (e.g., "The idea that you should be punished for your sins," "Feeling that you are watched or talked about by others"). Apparently, for common mental disorders, the generic applicability of the BSI is offset by a somewhat diminished ability to demonstrate change.

The hypothesis of a delayed response on the SF-36 was not supported by scores on the BSI and the SF-36. The data showed a similar linear pattern of change over time for the BSI total score and the mental component of the SF-36. We found a diminished response on the physical component score, but not a delayed response. Generally, scale scores of patients changed over time similarly, but to a lesser extent on physical component scales as compared to mental component scales and BSI scales. These results could have been affected by selective loss to follow-up, as four assessments were only available for 25% of the sample with at least two assessments. We compared the pretest scores of the sample with two assessments and the sample with four assessments and these were very similar. Furthermore, we inspected the course of scores over time of the samples with three ($n = 2946$) or two assessments ($n = 5786$) and found for these larger samples profiles that were very similar to Figs. 1 and 2. These results do not suggest that selective data loss explains the findings of the study.

When comparing the BSI with the SF-36 component scores, the physical component score demonstrated little change in this patient sample, and the BSI and the mental component score showed more or less equal amounts of change. This pattern of scores on the SF-36 is of course specific to patients treated for mental health problems. Likely, among patients treated for somatic diseases, the biggest change would occur on other SF-36 scales. For instance, Garratt et al. [44] compared change in score on the SF-36 scales over time for four groups of patients with somatic diseases (low back pain, menorrhagia, suspected peptic ulcer, and varicose veins). They found that the BP and RP scales revealed the largest changes. Likewise, ten Klooster et al. [45] demonstrated in patients with Rheumatoid Arthritis that the BP scale and the PCS showed the largest change in a retest period of six months. Finally, Frendl and Ware [46] reported on a meta-analysis of 185 drug trials in which they examined change on the component scores for fourteen different somatic conditions, and the PCS score showed overall slightly more change than the MCS. The PCS showed the largest change with psoriatic arthritis and rheumatoid arthritis; the MCS showed larger changes in depression and psoriasis. Nevertheless, when treating psychiatric patients, the physical component score of the SF-36 is still informative, as somatic

symptoms are important in their own right and can be an important cause for psychological distress [47, 48].

Further research aimed at broadening the scope of treatment outcome in mental health care research is needed and should focus on other concepts with potential relevance, such as the recovery concept [49]. Alternatively, in the direction of greater specificity, disorder-specific measurement instruments may yield more precise information on treatment gains [50]. The present findings of greater changes on the BSI depression and BSI anxiety subscales—in spite of their brevity—lend support to this suggestion. Finally, the more recent development of item response theory-based computer adaptive tests (CAT), such as the PROMIS assessment battery [51], may prove fruitful for outcomes research, as it allows for a more efficient assessment, without diminishing reliability which is usually associated with brief questionnaires.

A strength of the present study is its use of longitudinal data collected in everyday clinical practice, which enhances the generalizability of the findings. The size of the dataset implies ample statistical power to find differences between the outcome domains. On the other hand, the use of data collected under real-life circumstances yields less experimenter control, resulting in varied assessment intervals and substantial loss of data over time. Consequently, it is likely that treatment outcomes in the present study are somewhat inflated by selective loss of retest data, as patients who finished treatment unsuccessfully may decline to be reassessed. However, for a head-to-head comparison of outcome measures, the present data are very suitable, especially as the availability of lengthy assessments trajectories—four repeated assessments for a substantial number of patients—allowed for the investigation of synchrony of change on outcome domains.

Regarding the concordance between the BSI and the SF-36, it should be noted that the correlation coefficients presented in Table 3 may be a conservative estimate of the actual concordance of the underlying concepts of the instruments. The reliability of the BSI and the SF-36 scales determines the upper limit of their correlation according to the formula $r_{\max} = \sqrt{r_{xx}r_{yy}}$ [52]. The correlation between two scales can be corrected for their unreliability [53] with the formula $r^* = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$ (r^* is the attenuated correlation, r_{xy} the correlation between the scales and r_{xx} and r_{yy} are the test–retest reliability coefficients of the scales). With $r = .82$ for the Dutch version of the BSI–TOT score [13] and $r = .80$ for the SF-36-MCS [54], the correlation between the BSI and the MCS would increase from $r = .61$ to $r^* = .75$ and for EWB from $r = .75$ to $r^* = .93$, indicating that the measured concepts are even more concordant than the unattenuated correlation coefficients of Table 3 reveal.

Finally, the present study focused on whether both instruments assessed change of similar size and pace. While the change appears to be of similar size and had a synchronous course, this head-to-head comparison leaves unclear whether highly similar (latent) variable(s) or dimensions were assessed. In line with this, further research is needed to reveal for which population groups and in which situations one instrument is more advantageous compared to another [55].

Conclusion

We found correspondence but also significant differences between the BSI and the SF-36: change according to the BSI was similar to the mental component score (and its scales) of SF-36, but patients changed less on the physical component score and scale when compared to the mental component scores. Generally, the BSI and the SF-36 demonstrated a comparable degree of change in groups of patients, and this change occurs in similar size and pace. However, the profile of scores yielded by the SF-36 offers a more complete and more detailed clinical picture of the problems of individual patients, due to the additional domain of physical health.

Thus, the findings illustrate that there is considerable overlap between what is measured with the BSI and the SF-36, but also that each instrument contributes specific information regarding benefits from treatment. The BSI and the mental health component of the SF-36 offer similar specific information on symptom reduction or mental health gains. But the SF-36 clearly measures a broader construct and change on the physical component (and its scales) diverges from change on the SF-36 mental component as well as from change on the BSI. Finally, if the current findings regarding the substantial correlation between the mental component score and the BSI would be replicated with patients who are treated for somatic problems, the mental component scores of the SF-36 could be used to capture concurrent changes in psychological health.

Declarations

Conflict of interest The authors declared that they have no conflict of interest. EdeB developed the Dutch version of the BSI, but transferred copyright to Pearson Test Research, Inc. and does not receive any compensation for use of the BSI in the Netherlands.

References

1. Waldrop, J., & McGuinness, T. (2017). Measurement-based care in psychiatry. *Journal of Psychosocial Nursing and Mental Health Services, 55*(11), 30–35. https://doi.org/10.3928/02793695-20170818-01

2. Lambert, M. J. (2010). *Prevention of treatment failure. The use measuring, monitoring, and feedback in clinical practice.* American Psychological Association.

3. Kilbourne, A. M., Beck, K., Spaeth-Rublee, B., Ramanuj, P., O'Brien, R. W., Tomoyasu, N., et al. (2018). Measuring and improving the quality of mental health care: A global perspective. *World Psychiatry, 17*(1), 30–38. https://doi.org/10.1002/wps.20482

4. Greenhalgh, J. (2009). The applications of PROs in clinical practice: What are they, do they work, and why? *Quality of Life Research, 18*(1), 115–123. https://doi.org/10.1007/s11136-008-9430-6

5. Baumhauer, J. F., & Bozic, K. J. (2016). Value-based healthcare: Patient-reported outcomes in clinical decision making. *Clinical Orthopaedics and Related Research, 474*(6), 1375–1378. https://doi.org/10.1007/s11999-016-4813-4

6. Porter, M. E., & Teisberg, E. O. (2006). *Redefining health care: Creating value-based competition on results.* Harvard Business Press.

7. Drake, R. E., & Deegan, P. E. (2009). Shared decision making is an ethical imperative. *Psychiatric Services, 60*(8), 1007–1012. https://doi.org/10.1176/ps.2009.60.8.1007

8. Huber, M., Knottnerus, J. A., Green, L., van der Horst, H., Jadad, A. R., Kromhout, D., et al. (2011). How should we define health? *British Medical Journal, 343*(7817), 235–237. https://doi.org/10.1136/bmj.d4163

9. Thornicroft, G., & Slade, M. (2014). New trends in assessing the outcomes of mental health interventions. *World Psychiatry, 13*(2), 118–124. https://doi.org/10.1002/wps.20114

10. Putera, I. (2017). Redefining health: Implication for value-based healthcare reform. *Cureus, 9*(3), e1067–e1067. https://doi.org/10.7759/cureus.1067

11. Gaebel, W. (2016). Value-based mental healthcare: The quality aspect. *European Psychiatry, 33*(S1), S55–S55. https://doi.org/10.1016/j.eurpsy.2016.01.932

12. Baggaley, M. R. (2020). Value-based healthcare in mental health services. *British Journal of Psychiatry Advances, 26*(4), 198–204. https://doi.org/10.1192/bja.2019.82

13. de Beurs, E., & Zitman, F. G. (2006). De Brief Symptom Inventory (BSI): De betrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90 [The Brief Symptom Inventory: Reliability and validity of a handy alternative for the SCL-90]. *Maandblad Geestelijke Volksgezondheid, 61*, 120–141.

14. Derogatis, L. R. (1975). *The Brief Symptom Inventory.* Clinical Psychometric Research.

15. Slade, M., Amering, M., Farkas, M., Hamilton, B., O'Hagan, M., Panther, G., et al. (2014). Uses and abuses of recovery: Implementing recovery-oriented practices in mental health systems. *World Psychiatry, 13*(1), 12–20. https://doi.org/10.1002/wps.20084

16. Bakas, T., McLennon, S. M., Carpenter, J. S., Buelow, J. M., Otte, J. L., Hanna, K. M., et al. (2012). Systematic review of health-related quality of life models. *Health and Quality of Life Outcomes, 10*(1), 134. https://doi.org/10.1186/1477-7525-10-134

17. Slade, M. (2010). Mental illness and well-being: The central importance of positive psychology and recovery approaches. *BMC Health Services Research, 10*(26), 1–14. https://doi.org/10.1186/1472-6963-10-26

18. Seligman, M. E., Rashid, T., & Parks, A. C. (2006). Positive psychotherapy. *American Psychologist, 61*(8), 774–788. https://doi.org/10.1037/0003-066X.61.8.774

19. Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55*(1), 5–14. https://doi.org/10.1037/0003-066X.55.1.5

20. Jeste, D. V., Palmer, B. W., Rettew, D. C., & Boardman, S. (2015). Positive psychiatry: Its time has come. *Journal of Clinical Psychiatry, 76*(6), 675–683. https://doi.org/10.4088/JCP.14nr09599

21. Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.

22. Aaronson, N. K., Muller, M., Cohen, P. D., Essink-Bot, M.-L., Fekkes, M., Sanderman, R., et al. (1998). Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *Journal of Clinical Epidemiology, 51*(11), 1055–1068. https://doi.org/10.1016/S0895-4356(98)00097-3

23. Meesters, P. D., Comijs, H. C., de Haan, L., Smit, J. H., Eikelenboom, P., Beekman, A. T., et al. (2011). Symptomatic remission and associated factors in a catchment area based population of older patients with schizophrenia. *Schizophrenia Research, 126*(1), 237–244. https://doi.org/10.1016/j.schres.2010.09.003

24. Kortrijk, H. E., Mulder, C. L., van der Gaag, M., & Wiersma, D. (2012). Symptomatic and functional remission and its associations with quality of life in patients with psychotic disorder in Assertive Community Treatment teams. *Comprehensive Psychiatry, 53*(8), 1174–1180. https://doi.org/10.1016/j.comppsych.2012.05.001

25. Novick, D., Haro, J. M., Suarez, D., Lambert, M., Lépine, J.-P., & Naber, D. (2007). Symptomatic remission in previously untreated patients with schizophrenia: 2-Year results from the SOHO study. *Psychopharmacology (Berlin), 191*(4), 1015–1022. https://doi.org/10.1007/s00213-007-0730-2

26. Kennedy, S. H., Eisfeld, B. S., & Cooke, R. G. (2001). Quality of life: An important dimension in assessing the treatment of depression? *Journal of Psychiatry and Neuroscience, 26*(Suppl 1), 23.

27. Rapaport, M. H., Clary, C., Fayyad, R., & Endicott, J. (2005). Quality-of-Life impairment in depressive and anxiety disorders. *American Journal of Psychiatry, 162*(6), 1171–1178. https://doi.org/10.1176/appi.ajp.162.6.1171

28. Olatunji, B. O., Cisler, J. M., & Tolin, D. F. (2007). Quality of life in the anxiety disorders: A meta-analytic review. *Clinical Psychology Review, 27*(5), 572–581. https://doi.org/10.1016/j.cpr.2007.01.015

29. Sofko, C. A., Currier, J. M., & Drescher, K. D. (2016). Prospective associations between changes in mental health symptoms and health-related quality of life in veterans seeking posttraumatic stress disorder residential treatment. *Anxiety Stress and Coping, 29*(6), 1–14. https://doi.org/10.1080/10615806.2016.1157171

30. Cuijpers, P. (2019). Targets and outcomes of psychotherapies for mental disorders: An overview. *World Psychiatry, 18*(3), 276–285. https://doi.org/10.1002/wps.20661

31. Revicki, D. A., Kleinman, L., & Cella, D. (2014). A history of health-related quality of life outcomes in psychiatry. *Dialogues in Clinical Neuroscience, 16*(2), 127–135. https://doi.org/10.31887/DCNS.2014.16.2/drevicki

32. Iancu, S. C., Batelaan, N. M., Zweekhorst, M. B. M., Bunders, J. F. G., Veltman, D. J., Penninx, B. W. J. H., et al. (2014). Trajectories of functioning after remission from anxiety disorders: 2-Year course and outcome predictors. *Psychological Medicine, 44*(3), 593–605. https://doi.org/10.1017/S0033291713001050

33. Robinson, L., Delgadillo, J., & Kellett, S. (2020). The dose–response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research, 30*(1), 79–96. https://doi.org/10.1080/10503307.2019.1566676

34. de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., et al. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology and Psychotherapy, 18*(1), 1–12. https://doi.org/10.1002/cpp.696

35. Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., et al. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry, 59*(Suppl 20), 22–33. https://doi.org/10.1016/S0924-9338(97)83296-8

36. van Vliet, I. M., & de Beurs, E. (2007). *Tijdschrift voor Psychiatrie, 49*(6), 393–397.

37. van Fenema, E., Giltay, E., van Noorden, M., van Hemert, A., & Zitman, F. (2015). Assessing adherence to guidelines with administrative data in psychiatric outpatients. *Journal of Evaluation in Clinical Practice.* https://doi.org/10.1111/jep.12414

38. Ware, J. E., Gandek, B., Kosinski, M., Aaronson, N. K., Apolone, G., Brazier, J., et al. (1998). The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in 10 countries: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51*(11), 1167–1170. https://doi.org/10.1016/S0895-4356(98)00108-5

39. Ware, J. E., Koskinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A user's manual* (2nd ed.). The Health Institute.

40. Seidel, J. A., Miller, S. D., & Chow, D. L. (2013). Effect size calculations for the clinician: Methods and comparability. *Psychotherapy Research, 24*(4), 470–484. https://doi.org/10.1080/10503307.2013.840812

41. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

42. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

43. Sullivan, G. M., & Feinn, R. (2012). Using effect size—Or why the p value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

44. Garratt, A. M., Ruta, D. A., Abdalla, M. I., & Russell, I. T. (1994). SF 36 health survey questionnaire: II. Responsiveness to changes in health status in four common clinical conditions. *Quality in Health Care, 3*(4), 186–192. https://doi.org/10.1136/qshc.3.4.186

45. ten Klooster, P. M., Vonkeman, H. E., Taal, E., Siemons, L., Hendriks, L., de Jong, A. J. L., et al. (2013). Performance of the Dutch SF-36 version 2 as a measure of health-related quality of life in

patients with rheumatoid arthritis. *Health and Quality of Life Outcomes, 11*(1), 1–9. https://doi.org/10.1186/1477-7525-11-77

46. Frendl, D. M., & Ware, J. E. J. (2014). Patient-reported functional health and well-being outcomes with drug therapy: A systematic review of randomized trials using the SF-36 Health Survey. *Medical Care, 52*(5), 439–445. https://doi.org/10.1097/mlr.00000 00000000010311

47. Trivedi, M. H. (2004). The link between depression and physical symptoms. *Primary Care Companion to the Journal of Clinical Psychiatry, 6*(Suppl 1), 12–16.

48. Haug, T. T., Mykletun, A., & Dahl, A. A. (2004). The association between anxiety, depression, and somatic symptoms in a large population: The HUNT-II study. *Psychosomatic Medicine, 66*(6), 845–851. https://doi.org/10.1097/01.psy.0000145823.85658.0c

49. Anthony, W. A. (1993). Recovery from mental illness: The guiding vision of the mental health service system in the 1990s. *Psychosocial Rehabilitation Journal, 16*(4), 11–23. https://doi.org/10.1037/h0095655

50. Patrick, D. L., & Deyo, R. A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care, 27*(Suppl 3), 217–232. https://doi.org/10.1097/00005650-198903001-00018

51. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

52. Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *Journal of Experimental Education, 74*(3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266

53. Nunnally, J. C., & Bernstein, I. R. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

54. Ware, J. E., & Gandek, B. (1998). Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *Journal of Clinical Epidemiology, 51*(11), 903–912. https://doi.org/10.1016/S0895-4356(98)00081-X

55. Böhnke, J. R., & Croudace, T. J. (2016). Calibrating well-being, quality of life and common mental disorder items: Psychometric epidemiology in public mental health research. *British Journal of Psychiatry, 209*(2), 162–168. https://doi.org/10.1192/bjp.bp.115.165530