Received: 14 November 2017 | Revised: 3 April 2018 | Accepted: 29 May 2018

DOI: 10.1002/da.22809

RESEARCH ARTICLE

WILEY ADAA ANXIETY AND DEPRESSION ASSOCIATION OF AMERICA

Comparative responsiveness of generic versus disorder-specific instruments for depression: An assessment in three longitudinal datasets

Edwin de Beurs[1] | Ellen Vissers[2] | Robert Schoevers[2] | Ingrid V. E. Carlier[3] | Albert M. van Hemert[3] | Ybe Meesters[2]

[1] Faculty of Clinical Psychology, Leiden University, Leiden, The Netherlands

[2] Department of Psychiatry, University Medical Center Groningen, Groningen, The Netherlands

[3] Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
Edwin de Beurs, Faculty of Clinical Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, the Netherlands.
Email: e.de.beurs@fsw.leidenuniv.nl

Background: Routine outcome monitoring (ROM) may enhance individual treatment and is also advocated as a means to compare the outcome of different treatment programs or providers. There is debate on the optimal instruments to be used for these separate tasks.

Methods: Three sets with longitudinal data from ROM were analyzed with correlational analysis and repeated measures ANOVAs, allowing for a head-to-head comparison of measures regarding their sensitivity to detect change. The responsiveness of three disorder-specific instruments, the Beck Depression Inventory, the Inventory of Depressive Symptoms, and the Mood and Anxiety Symptoms Questionnaire, was compared to three generic instruments, the Symptom Checklist (SCL-90), the Outcome Questionnaire (OQ-45), and the Brief Symptom Inventory, respectively.

Results: In two of the three datasets, disorder-specific measures were more responsive compared to the total score on generic instruments. Subscale scores for depression embedded within generic instruments are second best and almost match disorder-specific scales in responsiveness. No evidence of a desynchronous response on outcome measures was found.

Limitations: The present study compares measures head-to-had, and responsiveness is not assessed against an external criterion, such as clinical recovery.

Discussion: Disorder-specific measures yield the most precise assessment for individual treatment and are recommended for clinical use. Generic measures may allow for comparisons across diagnostic groups and their embedded subscales approach the responsiveness of disorder-specific measures.

KEYWORDS
assessment, depression, psychometrics, responsiveness, routine outcome monitoring

1 | INTRODUCTION

There is evidence that routine outcome monitoring (ROM) has a positive influence on treatment efficacy (Lambert, 2010) and efficiency (Delgadillo et al., 2017) in mental health care. Routine measurements can be used in individual treatments by clinicians and patients to determine progress and the necessity to adapt interventions (Van, Dekker, Peen, van Aalst, & Schoevers, 2008), but they may also be used to evaluate the overall effectiveness of treatment programs, both within the same patient group as across diagnoses. The second goal could also include benchmarking, comparing outcomes of different teams or providers, in order to learn from each other and to improve treatment modalities (Barendregt, 2015). In the Netherlands, ROM has been implemented in mental healthcare with the aim to improve outcomes on the level of individual treatments as well as on the level of teams or providers (de Beurs, Barendregt, & Warmerdam, 2017). In terms of instruments, there is debate on whether the same instrument can be used for both goals, and which goal should be prioritized when time and effort are limited (Meesters, Duijzer, Nolen, Schoevers, & Ruhé, 2016).

Among measures to assess treatment outcome of depression, a distinction can be made between disorder-specific instruments, such as the Beck Depression Inventory (BDI) or the Inventory of

Depressive Symptomatology (IDS), and generic instruments for general psychopathology, such as the Outcome Questionnaire (OQ-45; Lambert, Gregersen, & Burlingame, 2004) or the Symptom Checklist/Brief Symptom Inventory (SCL-90/BSI; Derogatis, 1975a, 1975b). Advantage of generic instruments is that they allow for a comparison of treatment outcomes across diagnoses. Furthermore, they are convenient for use in everyday clinical practice, as the same instrument can be used to assess all patients. However, generic instruments may be less informative regarding specific symptoms of depression and less responsive to change as compared to instruments specifically designed to assess the severity of depression (Meesters et al., 2016).

The literature on the comparative responsiveness of generic and disorder-specific measures is expansive, but results are far from conclusive (Beaton, Bombardier, Katz, & Wright, 2001; Husted, Cook, Farewell, & Gladman, 2000; Terwee et al., 2007). Some studies report greater responsiveness of disorder-specific measures (Husted et al., 2000; Reine et al., 2005; Wiebe, Guyatt, Weaver, Matijevic, & Sidwell, 2003); others do not find a difference (Ades, Lu, & Madan, 2013; McCrindle et al., 2014; Tu, Hwang, Hsu, & Ma, 2017). The findings depend on the measures that are compared, the concepts assessed (e.g., quality of life vs. symptoms), the patient population under investigation, and the statistical approach chosen to investigate responsiveness (Terwee, Dekker, Wiersinga, Prummel, & Bossuyt, 2003). Several studies compare generic and disorder-specific measures of health-related quality of life (Ades et al., 2013; Wiebe et al., 2003), but comparisons between measures of clinical symptoms of mental disorders are less common (Hansson, Chotai, Nordstöm, & Bodlund, 2009; Wahl et al., 2014).

In psychiatry, research comparing responsiveness of generic and disorder-specific instruments for the assessment of the severity of psychopathology, is rather scarce. Most research compares quality of life measures, such as the EQ-5D and the SF-36 (Brazier et al., 2014), or is limited to cross-sectional analysis (Mauriño, Cordero, & Ballesteros, 2012). In the Netherlands, a study with eating disorder patients found superior responsiveness for the disorder-specific Eating Disorders Examination Questionnaire (EDE-Q; Fairburn & Beglin, 2008) over the generic BSI. The pre-to-posttest change was substantially larger on the EDE-Q ($ES_{EDE\text{-}Q} = 0.84$) than on the BSI ($ES_{BSI} = 0.45$). According to the BSI, 54.2% of the patients had improved or recovered according to the EDE-Q compared to 45.3% (Dingemans & van Furth, 2017). Recently, van der Mheen and colleagues compared various disorder-specific measures for anxiety with the generic OQ-45 and BSI and found superior responsiveness for disorder-specific measures as compared to the OQ-45 total score, but not compared to the BSI total score (van der Mheen, ter Mors, van den Hout, & Cath, 2018).

Responsiveness can be investigated by comparing results of different studies using distinct measures (Kounali, Button, Lewis, & Ades, 2016), but a stronger design is a "head-to-head" comparison of instrument in the same study (Wiebe et al., 2003). Responsiveness is investigated with correlational analyses or by comparing effect sizes (ES) of pre-to-posttest change according to various outcome measures (de Beurs et al., 2012; Husted et al., 2000). Responsiveness is best assessed and compared by investigating the course of scores over time of measures from the same longitudinal dataset. Although two assessments suffice to detect a difference in responsiveness, additional assessments allow for a more fine-grained analysis, for instance, detecting desynchrony of response over time (patients may change first on one measure and in a later phase of their treatment on another measure).

For an ongoing benchmark project in the Netherlands by Stichting Benchmark GGZ, the use of generic outcome scales is prescribed, such as the BSI (Derogatis, 1975a), the OQ-45 (Lambert et al., 2004), and the recently developed Dutch Symptoms Questionnaire-48 (Carlier et al., 2012; Carlier et al., 2017), to assess the outcome of treatment of psychiatric patients. However, a debate arose about the suitability of generic outcome scales (Meesters et al., 2016), as disorder-specific instruments, such as the BDI (Beck & Steer, 1990), the Inventory of Depressive Symptoms Self-report (IDS-SR; Rush, Gullion, Basco, Jarrett, & Trivedi, 1996), and the Mood and Anxiety Symptoms Questionnaire (MASQ; Watson et al., 1995), are possibly more responsive. Consequently, changes in symptomatology within patients may go unnoticed and we might get a too conservative estimate of the benefits of mental health care. Also, differences in performance between various providers may remain obscured when relatively unresponsive outcome measures are used.

In this paper, we report on three comparisons of generic and disorder-specific measures for depression in three separate datasets. We compared (1) BDI with SCL-90, (2) IDS-SR with OQ-45, and (3) BDI, MASQ, and BSI. We compared total scores on these instruments, but also subscale scores. Our hypothesis is that responsiveness improves with increased specificity for depression of the (sub)scale. We will also explore the data for signs of desynchrony in change over time between (sub)scales of the investigated instruments, as disorder-specific scales may detect change at an earlier phase of treatment than generic ones.

2 | METHODS

2.1 | Source of the data

Two datasets stem from the University Center Psychiatry (UCP) of the University Medical Center Groningen (UMCG), the Netherlands, a tertiary center for the treatment of mood, anxiety, and somatoform disorders. UCP provided data for the comparison of the BDI with the SCL-90 (collected from January 2012 until May 2016) and the IDS-SR with the OQ-45 (collected from September 2012 until May 2016). The Department of Psychiatry of the Leiden University Medical Center (LUMC; Centrum Onderzoek Routine Outcome Monitoring [COROM]) and the Mental Health Care Institute GGZ-Rivierduinen, providing secondary mental health care in the densely populated west of the country, yielded data for the BDI, MASQ, and BSI comparison (collected from April 2002 until October 2011). All patients from Leiden were diagnosed with the Mini International Neuropsychiatric Interview Plus, a structured diagnostic interview (MINI-plus; Sheehan et al., 1998; van Vliet & de Beurs, 2007). In Groningen, diagnoses (DSM IV) were determined in an intake session by a clinician.

All data were collected by ROM (de Beurs et al., 2011). Patients in ambulatory care, mostly seen weekly for psychological treatment and/or pharmacological treatment, were assessed at fixed intervals during their treatment, usually every 3 to 4 months. As treatments vary in duration, the number of assessments per patient varies as well. Data of instruments were matched on ID of the respondent and assessment moment (a match was declared when the assessment of the instruments had taken place in the same week, which was usually at the same occasion). The first dataset yielded longitudinal data with minimally two and a maximum of five consecutive assessments (N = 233); the second and third yielded a maximum of 10 assessments (N = 832 and N = 3,409, respectively). Patients gave permission for the (anonymized) use of their data for scientific purposes.

2.2 | Measures

2.2.1 | Disorder specific

The BDI-II (Beck & Steer, 1990) is a revised version of the original BDI (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and the most commonly used self-report instruments to assess the severity of depression. The questionnaire includes 21 items and each item has a set of four unique response options (0–3). For the Dutch translation, the BDI total score (BDI-TOT) and three subscale scores can be computed: the BDI cognitive factor (BDI-COG) with seven items, the BDI affective factor (BDI-AFF) with five items, and the BDI somatic factor (BDI-SOM) with nine items (van der Does, 2002).

The IDS-SR (Rush et al., 1996) also measures the severity of depression and includes 30 items with unique response options on a 4-point scale. A total score for 28 items (IDS-TOT) and two subscale scores for mood/cognition (IDS-MOOD) with 11 items and anxiety/arousal (IDS-ANX) with eight items (Wardenaar et al., 2010).

The MASQ (Watson et al., 1995) was developed to assess the symptoms of depression and anxiety disorders more distinctively. Items describe symptoms shortly and all have the same five response options on a Likert scale (not at all to very much). A total score (MASQ-TOT, 90 items) and subscale scores can be calculated for (lack of) positive affect (MASQ-PA; 22 items and specific for mood disorders), Somatic Anxiety (MASQ-SA; 18 items and specific for anxiety disorders) and negative affect (MASQ-NA; 20 items and nonspecific to depression or anxiety). The factorial structure is replicated in the Dutch translation of the MASQ by de Beurs, den Hollander-Gijsman, Helmich, and Zitman (2007).

All these instruments have good psychometric properties, both the U.S. originals and their Dutch translations (de Beurs et al., 2007; van der Does, 2002; Wardenaar et al., 2010).

2.2.2 | Generic specific

The SCL-90 (Derogatis, 1975b) is a generic self-report questionnaire for the severity of psychopathology (Koeter, Ormel, & van den Brink, 1988). The checklist includes 90 short descriptions of problems or symptoms, and respondents are asked to indicate how much they were bothered in the last week on a Likert scale with five response options (not at all to very much). The total score (SCL-TOT) is the sum of all responses with a range of 90–450. The Dutch translation has good psychometric properties (Arrindell & Ettema, 1986), but uses a different dimensional structure than the U.S. original with eight subscales: anxiety, agoraphobia, depression (SCL-DEP), somatization, insufficiency in thinking or behavior (SCL-IN), interpersonal sensitivity, hostility (SCL-HOS), and sleeping problems.

The BSI (Derogatis, 1975a) is a shortened version of the SCL-90 with 53 items with identical instructions and response options. The total score on the 53 items is the BSI-TOT and is calculated as the mean score on all items with a range of 0–4. The Dutch BSI has nine subscales—in accordance with the U.S. original—of each five to six items: depression (BSI-DEP), somatization (BSI-SOM), anxiety, phobia, interpersonal sensitivity, obsessive compulsive, hostility, paranoia, and psychoticism. The Dutch version of the BSI has good psychometric properties (de Beurs & Zitman, 2006).

The OQ-45 (Lambert et al., 2004) has 45 items, asking the patient to indicate how often a symptom or problem occurred in the week prior to the assessment on a Likert scale (never–almost always). A total of 25 items describe problems and symptoms and compose the symptomatic distress scale (OQ-SD, range 0–100); 11 items assess functioning with family and friends, and compose interpersonal relations scale (OQ-IR, range 0–44); and nine items assess functioning at work or school, and compose the social role scale (OQ-SR, range 0–36). Also, a total score can be computed for general well-being (OQ-TOT, range 0–180). The Dutch translation has good psychometric properties (de Beurs, den Hollander-Gijsman, Buwalda, Trijsburg, & Zitman, 2005; de Jong et al., 2007). On all instruments, higher scores indicate more severe depression, more symptoms or problems, and worse functioning.

The dichotomy of generic and disorder-specific does not hold when subscale scores on generic instruments are also taken into consideration. Table 1 shows a proposal on how to place instruments and their subscales on the dimension generic–depression specific. The OQ-TOT is deemed the most generic as this score includes functioning as well as symptomatology. Next, the OQ-SD and the SCL-90/BSI-TOT assess general symptomatology/psychopathology. The MASQ-TOT is somewhat more specific, assessing only symptoms of mood and anxiety disorders. Negative affect is specific to depression at the same level as the BDI-TOT and IDS-TOT. Finally, the IDS-SR and BDI-subscales and the MASQ-PA are deemed to be the most specific for depression.

2.3 | Statistical analyses

To put all instruments on a common metric, scores were standardized on the pretest mean and standard deviation. Consequently, the entire population gets a baseline score of $M = 0$ ($SD = 1$) on all scales. Scores diminish over time and the absolute value of mean score at consecutive assessments represents the within groups effect size (*ES*) for each (sub)scale. We established the *ES* for the first and second assessment (initial treatment phase) and for the first and last available (*n*th) assessment (the maximum pre-to-posttest change).

In each dataset, the correlations between (sub)scales of instruments at baseline and at the last available assessment were calculated, Furthermore, the correlations between difference scores for the first assessment interval and for the maximum interval (first to *n*th assessment) were calculated. High correlations between differences scores suggest similar responsiveness.

Responsiveness was also investigated with repeated measures ANOVAs on subsets of the three samples with complete data for three,

TABLE 1 Position of total scale scores and subscale scores of the instruments on the dimension from generic to disorder-specific

	Dimension:				
Scale:	**Generic**				**Specific**
OQ	OQ-TOT	OQ-SD		IDS-TOT	IDS-MOOD
BDI				BDI-TOT	BDI-AFF/COG
MASQ			MASQ-TOT	MASQ-NA	MASQ-PA
SCL/BSI		SCL/BSI-TOT		SCL/BSI-DEP	

Notes. OQ-TOT, OQ-45 total score; OQ-SD, symptomatic distress; OQ-IR, interpersonal relations; IDS-TOT, IDS-SR total score; IDS-MOOD, mood symptoms; BDI-TOT, BDI total score; BDI-AFF, affective factor; BDI-COG, cognitive factor; SCL/BSI-TOT, SCL-90 or BSI total score; SCL/BSI-DEP, SCL-90 or BSI depression; MASQ-TOT, MASQ total score; MASQ-NA, negative affect; MASQ-PA, (lack of) positive affect.

four, or five assessments. For instance, as the first dataset yielded sufficient subjects with at least three assessments, we compared their total score on the SCL-90 with their total score on the BDI with a repeated measures ANOVA in a 2 (instrument) × 3 (time) factorial design (see Table 3); in a similar vein, we compared the SCL-90 depression subscale score with the BDI-TOT. These two ANOVAs were repeated for the maximum treatment effect in a 2 (instrument) × 2 (time; the first and the *n*th assessment) design. All analyses yield a time effect (scores change over time irrespective of the instrument), an instrument effect (standardized scores differ between instruments irrespective of time), and a time-by-instrument interaction. This interaction effect is most informative, as it signifies a difference in responsiveness between instruments: the score on instrument A decreases more over time (or sooner in time) than the score on instrument B. The analyses yield an *F*-statistic, *P*-value, and effect size indicator (η^2) for each effect. As the size of the datasets yields ample power to find statistically significant differences over time or between measures, the ES of the ANOVAs are most informative; η^2 can be interpreted as the amount of variance explained and—according to the rule of thumb of Cohen (1988)—$\eta^2 = 0.020$ is a small effect, $\eta^2 = 0.130$ is an intermediate effect, and $\eta^2 = 0.260$ is a large effect.

3 | RESULTS

3.1 | The three datasets

In the dataset for the comparison of the BDI and the SCL-90 ($N = 233$), the mean age of respondents was $M = 33.7$ years ($SD = 11.1$) and 52.8% was female. All patients suffered from a mood disorder, although this was not for all the primary diagnosis. According to their primary diagnosis, 29.6% of patients had a mood disorder (major depressive disorder or dysthymia), 34.8% had a somatoform disorder, 18.5% had an anxiety disorder, and 17.2% had another diagnosis. In the dataset comparing the IDS-SR and the OQ-45 ($N = 832$), mean age of patients was $M = 38.4$ years ($SD = 13.7$) and 54.9% was female; here, 60.3% of patients had a primary diagnosis of mood disorder, 27.1% had an anxiety disorder, 3.7% had a somatoform disorder, and 8.9% had another primary diagnosis. Both datasets comprised patients from the UCP of the UMCG and represent patients typically seen in an outpatient clinic for common mental disorders. In the third dataset, stemming from Leiden, the BSI–BDI-MASQ dataset ($N = 3409$), the mean age of patients was $M = 40.7$ years ($SD = 13.4$) and 64.0% was female. All patients were referred for treatment for mood, anxiety, or somatoform disorder and all were diagnosed with the MINI-plus (Sheehan et al., 1998). Here, no distinction was made between primary and secondary disorders. A total of 2,370 patients (69.5%) met criteria of a (single or comorbid) mood disorder; 340 (10.0%) had an anxiety, or somatoform disorder without a comorbid depression; 699 (20.5%) did not meet formal criteria for a current mood, anxiety, or somatoform disorder. In all three datasets, data were reanalyzed in a pure subset of patients with a primary mood disorder diagnosis ($n = 69$ and 502) or a mood disorder according to the MINI-plus ($n = 2{,}370$).

Table 2 presents the number of patients per assessment and the average assessment interval. The mean length of the total assessment interval was $M = 371$ days ($SD = 183$) for the BDI–SCL-90 comparison, $M = 394$ days ($SD = 282$) for the IDS–OQ-45 pair, and $M = 315$ days ($SD = 250$) for the BSI–BDI-MASQ comparison.

3.2 | Comparison of BDI and SCL-90

Supporting Information Table A shows Pearson correlation coefficients between BDI and SCL-90 (sub)scales at baseline, at the last assessment, and for the difference score for the first assessment interval (and for the maximal difference score, the first to *n*th assessment). The correlations are of medium or high size (>0.70); correlations between the BDI-TOT, the SCL-TOT, and the SCL-DEP are the highest. This applies to the cross-sectional correlations, but also for the longitudinal difference scores. At first glance, these scales measure the same concept, and difference scores are also substantially associated (Supporting Information Table B).

Figure 1 (left) shows the maximum ES for the (sub)scales of the BDI and the SCL-90, ordered from high to low. The BDI-TOT and BDI-COG are the most responsive ($ES_{\text{BDI-TOT}} = 0.96$ and $ES_{\text{BDI-COG}} = 0.91$); the SCL-TOT is less responsive ($ES_{\text{SCL-TOT}} = 0.75$); and the SCL-DEP is somewhat more responsive ($ES_{\text{SCL-DEP}} = 0.82$). Other subscales of the SCL-90 are again less responsive ($ES_{\text{SCL-ANG}} = 0.62$ and $ES_{\text{SCL-HOS}} = 0.37$).

For this dataset, 160 patients had complete data at three assessments and their data were analyzed with repeated measures ANOVAs. The results are presented in Table 3 and show a large time effect ($\eta^2 = 0.413$). Irrespective of the instrument, scores diminish and patients improve. There is also a small to intermediate instrument effect and a small to intermediate interaction effect: the responsiveness of instruments differs. The results in the Supporting Information Table B, which displays more detailed information on the significant effects using a repeated time and a simple instrument contrast in the

TABLE 2 Number of patients and length of time intervals in days for the full samples and for the "depression-only" subsamples

BDI and SCL-90	First and second	Third	Fourth	Fifth assessment
N	233	160	48	6
Interval: *M* (*SD*)	146.5 (62.8)	243.8 (121.8)	253.7 (71.4)	195.8 (80.6)
Subset with depression only				
N	69	52	23	1
Interval: *M* (*SD*)	151.0 (43.7)	202.7 (97.1)	268.7 (58.6)	76.0
IDS-SR and OQ-45				
N	832	402	210	107
Interval: *M* (*SD*)	222.8 (191.2)	168.5 (134.0)	165.9 (122.3)	194.5 (150.2)
Subset with depression only				
N	502	238	132	79
Interval: *M* (*SD*)	204.5 (179.4)	168.4 (141.2)	158.1 (123.9)	198.2 (145.1)
BDI, BSI, and MASQ				
N	3409	1804	943	479
Interval: *M* (*SD*)	149.0 (63.4)	140.9 (60.5)	141.0 (59.0)	147.9 (61.7)
Subset with depression only				
N	2370	1311	705	370
Interval: *M* (*SD*)	150.7 (64.8)	140.5 (60.9)	141.5 (58.3)	146.8 (60.8)

Notes. BDI, beck depression inventory; SCL-90, symptom checklist; IDS-SR, inventory of depressive symptoms self-report; OQ-45, outcome questionnaire; BSI, brief symptom inventory; MASQ, mood and anxiety symptoms questionnaire.

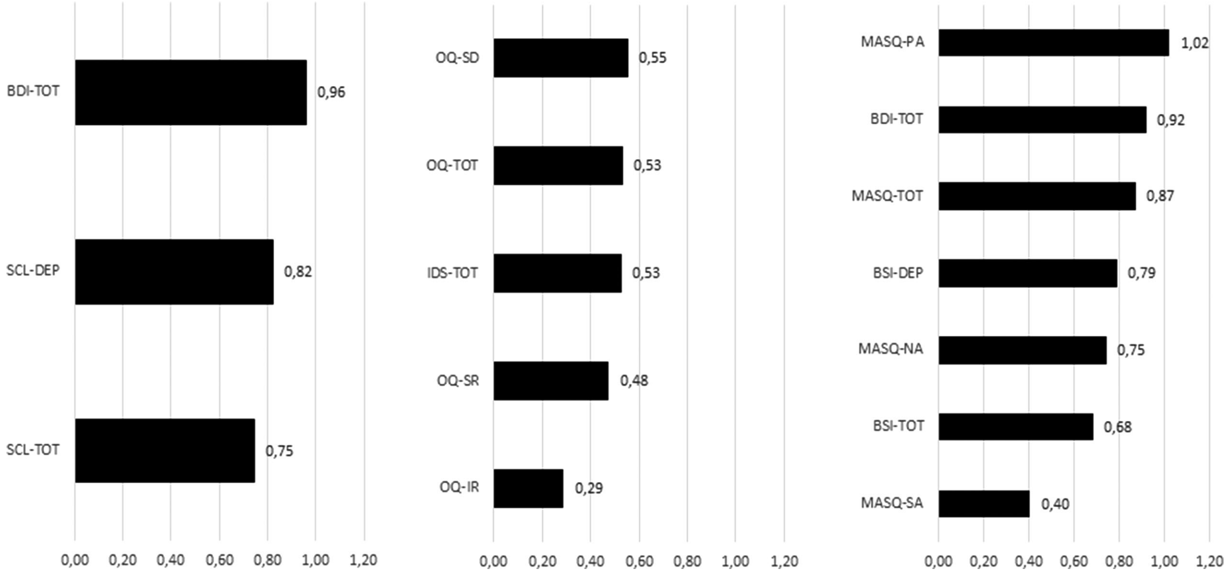

FIGURE 1 Responsiveness (standardized ES) of (sub)scales in the three datasets: the SCL-90 and BDI scales (left), the IDS-SR and OQ-45 (middle), and the BSI, MASQ, and BDI scales (right), rank ordered by size

TABLE 3 Overview of results of repeated measures ANOVAs for various comparisons of BDI and SCL (sub)scales

			Time			Instrument			Time × Instrument^2		
Comparison		*N*	*df*	*F*	η^2	*df*	*F*	η^2	*df*	*F*	η^2
BDI vs. SCL-TOT	First, second, and third	160	2,318	111.64	**0.413**	1,159	28.19	**0.151**	2,318	14.54	**0.087**
BDI vs. SCL-DEP	First, second, and third	160	2,318	93.69	**0.397**	1,159	18.94	**0.011**	2,318	9.47	**0.056**
BDI vs. SCL-TOT	First vs. *n*th	233	1,232	181.50	**0.439**	1,232	12.15	**0.050**	1,232	20.44	**0.081**
BDI vs. SCL-DEP	First vs. *n*th	233	1,232	178.77	**0.435**	1,232	7.89	**0.033**	1,232	7.49	**0.031**

Notes. BDI, BDI total score; SCL-TOT, SCL-90 total score; SCL-DEP, SCL-90 Depression subscale; statistically significant main and interaction effects ($P < 0.05$) are indicated by η^2 in bold typeface.

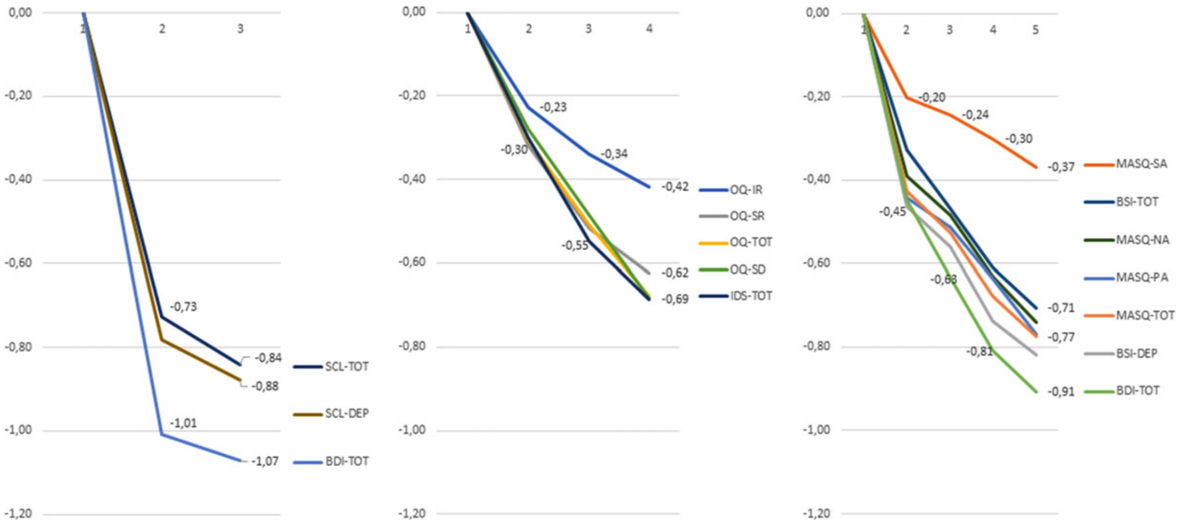

FIGURE 2 Course over time of standardized scores on the SCL90 and BDI for the group with three assessments ($n = 160$); on the IDS and OQ for the subgroup with four assessments ($n = 210$), and on the BDI, BSI, and MASQ for the group with five assessments ($n = 479$)

ANOVAs, show that the time effect is the largest in the initial phase of the treatment ($\eta^2 = 0.442$. vs. $\eta^2 = 0.023$ for the first and the second phase of the treatment, respectively). Furthermore, the instrument contrast (a simple contrast comparing the BDI-COG with the other scales) reveals that the BDI-COG factor is more responsive than most of the SCL-90 subscales except for the SCL-DEP and SCL-IN scales. A further distinction by therapy phase indicates that the instrument-by-time interactions are only significant in the first phase (again except for the SCL-DEP and SCL-IN). The course of scores over time is shown in Figure 2 (left). It illustrates that the BDI-TOT is the most responsive scale.

3.3 | Comparison of IDS-SR and OQ-45

Supporting Information Table C presents Pearson correlation coefficient between the IDS-SR and the OQ-45 and their subscales. The correlation between the IDS-SR scales and the OQ-45 OQ-SD are high; correlations are lower between the IDS-SR and the functioning scales of the OQ-45 (OQ-SR and OQ-IR).

Figure 1 (middle) shows the maximum ES values from $N = 832$ respondents. The OQ-SD has the largest pre–post difference ($ES_{OQ\text{-}SD} = 0.55$), followed by the OQ-TOT, the IDS-TOT, and the OQ-SR ($ES_{OQ\text{-}SR} = 0.48$). The OQ-IR is less responsive.

A subset of these data was analyzed with repeated measures ANOVAs. As Table 4 shows, only the analyses comparing subscales of the instruments yield a significant (but small) interaction effect. The contrasts in Supporting Information Table D show that also in this dataset, the largest decrease in scores occurs in the first months of treatment ($\eta^2 = 0.121$, 0.055, and 0.029 in the first three phases of treatment). The only significant interaction effect is the difference in responsiveness between the IDS-MOOD and the OQ-IR subscale during the second treatment phase ($\eta^2 = 0.24$).

The course of scores over at least four assessments is depicted in Figure 2 (middle). Four subscales indicate a similar decline over time: IDS-TOT, OQ-SD, OQ-TOT, and OQ-SR (ES ranges from 0.62 to 0.69). Again, the OQ-IR is less responsive to change, with $ES_{OQ\text{-}IR} = 0.42$.

3.4 | Comparison of BDI, MASQ, and BSI

Supporting Information Table E presents correlation coefficients between the BDI and the BSI; Supporting Information Table F shows correlation coefficients between the BDI and MASQ and the BSI and MASQ. Scales measuring corresponding constructs correlate considerably, both cross-sectionally and longitudinally (difference scores). This holds for depression, but also for somatic anxiety (the MASQ-SA) and somatic symptoms (the BSI-SOM). Lack of positive affect of the MASQ (supposedly the hallmark of depression; cf. Clark, Watson, & Mineka, 1994) is less strongly correlated with depression according to the BDI or the BSI-DEP score. Figure 1 (right) shows maximum ES for the total scores and subscale scores of the BDI, MASQ, and BSI. The positive affect scale (the subscale of the MASQ specific for depression) appears the most responsive ($ES_{MASQ\text{-}PA} = 1.02$); the BDI-TOT, BDI-SOM, BDI-AFF, MASQ-TOT, and the BSI-DEP ($ES_{BSI\text{-}DEP} = 0.79$) are all still fairly responsive. The BSI subscales not assessing depression are the least responsive ($ES_{BSI\text{-}SOM} = 0.40$).

Table 5 shows the results of repeated measures ANOVAs. The scores of five assessments were used. There were significant differences in responsiveness between total and subscale scores (small to intermediate η^2), which is illustrated in Figure 1 (right): the MASQ-PA and the BDI-TOT scale were the most responsive ($ES_{MASQ\text{-}PA} = 1.02$ and $ES_{BDI\text{-}TOT} = 0.92$); the BSI-DEP and the BSI-TOT were less responsive ($ES_{BSI\text{-}DEP} = 0.79$ and $ES_{BSI\text{-}TOT} = 0.71$). Almost all time and instrument contrasts are statistically significant. ES (η^2) in Supporting Information Table G reveals that the largest differences in responsiveness are between the BDI-TOT (the most responsive scale in this dataset) and the BSI and MASQ subscales not measuring depression (e.g., MASQ-SA). The course over time is illustrated in Figure 2 (right).

Finally, all analyses were repeated for more homogenous diagnostic groups of only patients diagnosed with depression and patients diagnosed with singular depression (without comorbid conditions).

TABLE 4 Overview of results of repeated measures ANOVA of various comparisons of IDS and OQ-45 (sub)scales

			Time			Instrument			Time × instrument		
Comparison		***N***	***df***	***F***	η^2	***df***	***F***	η^2	***df***	***F***	η^2
IDS vs. OQ-TOT	First, second, third, and fourth	210	3,627	42.34	**0.168**	1,209	0.09	0.004	3,627	0.53	0.003
IDS vs. OQ-SD	First, second, third, and fourth	210	3,627	40.98	**0.164**	1,209	0.11	0.001	3,627	1.19	0.006
IDS vs. OQ-TOT	First vs. *n*th	832	1,831	252.10	**0.233**	1,831	1.05	0.001	1,831	0.02	0.000
IDS vs. OQ-SD	First vs. *n*th	832	1,831	254.08	**0.234**	1,831	1.78	0.003	1,831	1.81	0.002

Notes. IDS, IDS-SR total score; OQ-TOT, total score; OQ-subs, OQ-45 subscales; OQ-SD, symptomatic distress subscale; statistically significant main and interaction effects ($P < 0.05$) are indicated by η^2 in bold typeface.

TABLE 5 Overview of results of repeated measures ANOVA of various comparisons of BDI, MASQ, and BSI (subscales)

			Time			Instrument			Time × instrument		
		N	***df***	***F***	η^2	***df***	***F***	η^2	***df***	***F***	η^2
TOT scales	First to Fifth	479	4,1912	123.87	**0.206**	2,1912	37.81	**0.073**	8,3824	11.87	**0.024**
BDI, PA, BSI-DEP	First to Fifth	479	4,1912	109.42	**0.186**	2,956	0.77	0.002	8,3824	3.79	**0.008**
TOT scales	First vs. *n*th	3,409	1,3408	2470.88	**0.420**	2,3408	128.70	**0.036**	2,6816	262.19	**0.071**
BDI, PA, BSI-DEP	First vs. *n*th	3409	2,3408	2460.64	**0.419**	2,6816	82.84	**0.024**	2,6816	166.74	**0.047**

Notes. TOT-scales, total score on BDI, MASQ, and BSI; BDI, BDI-total score; PA, MASQ-positive affect; BSI-DEP, BSI-depression.

By and large, the findings of these analysis regarding the comparative responsiveness of the instruments mimicked the results found with the entire sample, as reported in Table 5 and Supporting Information Tables E, F, and G. Correlational analyses and ANOVAs with the mood disorder samples were similar to the earlier findings; the pre-to-posttest change on scales was somewhat larger in comparison to the diagnostically more heterogeneous samples (~0.10 ES points), but the comparative responsiveness remained the same.

4 | DISCUSSION

Correlational analyses revealed a pattern of associations among the measures and their subscales, which is in line with the concepts they intend to measure. Associations among (difference) scores between (sub)scales for the same construct are high and lower for distinct constructs. Correlations between difference scores tend to be lower compared to correlations between scores at single assessment occasions, due to inter-individual variation in change and due to the lower reliability of difference scores. Measurement errors at both time points limit the reliability of difference scores, which is by definition lower than the reliability of single scores (Cronbach, 1984). At first glance, especially the correlational results between difference scores suggest similar responsiveness of generic and disorder-specific instruments for depression.

Repeated measures ANOVA revealed differences in responsiveness in two of the three datasets in the expected direction: disorder-specific scales tend to be more responsive than total scores on generic scales, although the ES of the statistically significant time-by-instrument effects are small. Generally, no evidence of desynchrony in response was found. The most responsive scale is the BDI-TOT, but not its subscale scores. The BDI-TOT is 1.3 times more responsive than the SCL-90 total score in the first dataset and 1.4 times more responsive than the BSI-total score in the third dataset. Differences between the BDI-TOT and the depression subscales of the SCL-90 or BSI are somewhat less pronounced, but the BDI is still 1.2 times more responsive. The differences in responsiveness between the instruments may be due to various factors, as their psychometric properties diverge, such as reliability (test–retest and internal consistency, unidimensionality of the scales, and the number of items), properties of items (information value and scalability), and properties of response formats (generic vs. specifically adapted to the item content, such as the BDI). Finally, the validity of the scales (how well the scales cover the concept of depression) is important: depression scales are dedicated to the measurement of depression, whereas generic scales may miss some relevant aspects of depression and their total score includes items that are irrelevant to depression.

According to the ANOVA for repeated measurements, the depression subscales of the SCL-90 (Table 3 and Supporting Information Table B) and the BSI (Table 5 and Supporting Information Table G) do not differ from the most responsive subscale of the BDI in their ability to detect chance. Thus, our hypothesis that the most specific scales will be the most responsive is not fully confirmed by the results. Within the generic SCL-90 and BSI, subscales for depression are more responsive than total scores on these instruments, but within the disorder-specific BDI and IDS, subscales are generally not more responsive than the total scores. Finally, we see no desynchrony in responsiveness among the compared (sub)scales.

The three samples in this study were composed of patients with common mental disorders, predominantly depression, but also some patients with anxiety or somatoform disorders were included, who did not meet formal diagnostic criteria for a (comorbid) mood disorder. The findings should be interpreted with caution as the data were obtained in everyday clinical practice and stem from a somewhat heterogeneous sample. Repeating the analyses in the sample, after removing the patients without a depression diagnosis, did not alter the results in a meaningful way. Finally, we reanalyzed the data from patient with a singular mood disorder (which was only possible in the Leiden sample with complete diagnostic information). The results were highly similar to the initial findings, except for lower baseline scores of patients with a singular mood disorder and a slight increase in the pre-to-posttest change in the pure depression groups compared to the more diagnostically mixed samples. Apparently, both in the purer mood disorder samples as well as in the more heterogeneous samples, total scores on disorder specific scales are more responsive than total scores on generic scales. Most items of the BDI or IDS are relevant for this patient group, whereas in the SCL-90 or BSI, there are many items with less relevance, lowering the SCL90 or BSI-TOT at baseline and thus making these scales less responsive to change.

Surprisingly, the second dataset showed that the IDS-SR total score and its subscale scores are not more responsive than the OQ-45 or its subscales. A possible explanation for not finding this expected difference in responsiveness may be that in general less change in symptoms was realized in this sample, comprising relatively treatment-resistant patients, leaving also less room for distinctive responsiveness of both instruments. However, subscale scores on the IDS-SR appear even less responsive compared to subscale scores on the OQ-SD (see Table 4 and Figure 2 [middle]). The IDS was composed by Rush and colleagues (1986) to cover a broad range of symptoms in order to be more sensitive to change than the gold standard for depression assessment at that time, the Hamilton Depression Rating Scale (HDRS; Hamilton, 1960). There is a rating scale version to be completed by a clinician (IDS-C) and a self-report version (IDS-SF). Indeed, Rush et al. (1996) report superior responsiveness of the IDS-C over the HDRS. Helmreich and colleagues (2011) replicated this finding. However, the IDS-SR seems less responsive than the IDS-C, as Fried et al. (2016) have shown. Moreover, they raise doubt on the unidimensionality and measurement invariance over time of the IDS-SR. Finally, Corruble and colleagues (1999) compared the IDS, the Montgomery-Åsberg Depression Rating Scale (Montgomery & Asberg, 1979), and the SCL-90 in psychiatric inpatients and reported equal responsiveness of the IDS-SR and the SCL-DEP subscale, a finding in line with our current findings. The plea of Meesters et al. (2016) to replace the OQ-45 and use the IDS-SR instead for benchmarking in the Netherlands is not supported by the present findings.

Strength of the present study is that the available data offered a unique opportunity for a head-to-head comparison of generic and disorder-specific instruments. A further strength is the size of the three datasets, providing sufficient data to yield ample statistical power to find statistically significant differences in responsiveness between self-report instruments. Furthermore, these data were collected in real-life clinical practice with diverse samples of psychiatric patients (mild to moderate common mental disorders in ambulatory care in the Leiden sample and a mix of in- and outpatients with moderate to severe problems in the samples from Groningen). Finally, the analyses on three or more assessments allowed us to investigate not only differences in responsiveness among measures, but also to check for potential desynchrony in change over time.

5 | LIMITATIONS

Not all possible comparisons between measures were feasible. Unfortunately, a dataset for a head-to-head comparison of the IDS-SR with the SCL-90 or BSI was not available. The IDS-SR was only compared with the OQ-45, and this subsample showed the lowest ES for the pre-to-posttest change after four assessments, which limits the chance of finding differences in responsiveness between instruments in the first place. Future research should reveal how the IDS-SR and the BSI depression subscale compare in responsiveness. Furthermore, this study compared Dutch versions of self-report questionnaires and we recommend replication of this research with the original versions in English-speaking samples. Finally, all results and conclusions apply to the situation that we want optimal information from aggregated data, for example, when comparing outcomes of patient groups in a randomized controlled trial or in a naturalistic observational study. For ROM, where the focus is on monitoring progress of individual patients, specific information may be required, and this can best be provided by the administration of (subscales of) disorder-specific instruments. A final limitation of our study is that it merely focused on internal responsiveness or the ability of instruments to measure change over time. The value of the responsiveness index is dependent on the actual change achieved, which may diverge per study of per patient sample, as turned out with our three data sets. In contrast, external responsiveness attempts to denote responsiveness as the relationship between change in a measurement and change in an external standard (Husted et al., 2000). External responsiveness of an instrument may be expressed in how well it distinguishes between recovered and unchanged groups of patients. Future research may also evaluate the external responsiveness of generic and disorder-specific mental health measures. However, establishing external responsiveness does require an external criterion to decide on the clinical status of psychiatric patients, which is somewhat problematic in common mental disorders, where the transition of functional to dysfunctional is usually gradual. This is in line with a more dimensional approach toward conceptualizing psychopathology. Also, alternative views on the structure of psychopathology (Krueger & Markon, 2006; Walton, Ormel, & Krueger, 2011) and other models for the association among psychopathology symptoms and symptom clusters, such as network models, (Borsboom & Cramer, 2013; Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016) deserve attention in future development and research of outcome measures for depression treatment.

6 | CONCLUSION

In conclusion, the disorder-specific BDI is more responsive compared to the total score on the generic SCL-90, and BSI and the disorder-specific IDS does not appear more responsive than the OQ-45. The responsiveness of the depression subscales of the SCL-90/BSI falls in between. For an efficient assessment of symptomatology in a sample with diverse psychiatric disorders, it may be sufficient to administer a generic instrument, preferably with responsive subscales for specific problems. However, for optimum power to detect differences in a trial or for detailed information on individual patients, use of the more responsive disorder-specific instruments is recommended.

ACKNOWLEDGMENT

The essential contribution made by patients participating in ROM at the UMCG, the LUMC, and the mental health care provider GGZ Rivierduinen is very gratefully acknowledged.

ORCID

Edwin de Beurs http://orcid.org/0000-0003-3832-8477

REFERENCES

Ades, A., Lu, G., & Madan, J. J. (2013). Which health-related quality-of-life outcome when planning randomized trials: Disease-specific or generic, or both? A common factor model. *Value in Health*, *16*, 185–194.

Arrindell, W. A., & Ettema, J. H. M. (1986). *SCL-90: Handleiding bij een multidimensionele psychopathologie-indicator [SCL-90: Manual for a multidimensional indicator of psychopathology]*. Lisse, the Netherlands: Swets & Zeitlinger.

Barendregt, M. (2015). Benchmarken en andere functies van ROM: Back to basics [Benchmarking and other functions of ROM: Back to basics]. *Tijdschrift voor Psychiatrie*, *57*, 517–525.

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology*, *54*, 1204–1217.

Beck, A. T., & Steer, R. A. (1990). *Manual for the beck anxiety inventory*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Brazier, J., Connell, J., Papaioannou, D., Mukuria, C., Mulhern, B., Peasgood, T., … Barkham, M. (2014). A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technology Assessment*, *18*, 1–188.

Carlier, I., Schulte-Van Maaren, Y., Wardenaar, K., Giltay, E., Van Noorden, M., Vergeer, P., & Zitman, F. (2012). Development and validation of the 48-item Symptom Questionnaire (SQ-48) in patients with depressive, anxiety and somatoform disorders. *Psychiatry Research*, *200*, 904–910.

Carlier, I. V. E., Kovács, V., van Noorden, M. S., van der Feltz-Cornelis, C., Mooij, N., Schulte-van Maaren, Y. W. M., … Giltay, E. J. (2017). Evaluating the responsiveness to therapeutic change with Routine Outcome Monitoring: A comparison of the Symptom Questionnaire-48 (SQ-48) with the Brief Symptom Inventory (BSI) and the Outcome Questionnaire-45 (OQ-45). *Clinical Psychology & Psychotherapy*, *24*, 61–71. https://doi.org/10.1002/cpp.1978

Clark, L. A., Watson, D., & Mineka, S. (1994). Temperament, personality, and the mood and anxiety disorders. *Journal of Abnormal Psychology*, *103*, 103–116.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Corruble, E., Legrand, J., Duret, C., Charles, G., & Guelfi, J. (1999). IDS-C and IDS-sr: Psychometric properties in depressed in-patients. *Journal of Affective Disorders*, *56*, 95–101.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York, NY: Harper & Row.

de Beurs, E., Barendregt, M., Flens, G., van Dijk, E., Huijbrechts, I., & Meerding, J. W. (2012). Vooruitgang in de behandeling meten: Een vergelijking van vragenlijsten voor zelfrapportage [Measuring treatment progress: A comparison of self-report questionnaires for treatment outcome]. *Maandblad Geestelijke Volksgezondheid*, *67*, 259–270.

de Beurs, E., & Barendregt, M., & Warmerdam, L., (Eds.). (2017). *Behandeluitkomsten: Bron voor kwaliteitsbeleid in de GGZ [Treatment outcome: Source of quality management in mental Health Care]*. Amsterdam, the Netherlands: Boom.

de Beurs, E., den Hollander-Gijsman, M., Buwalda, V., Trijsburg, W., & Zitman, F. G. (2005). De Outcome Questionnaire (OQ-45): Een meetinstrument voor meer dan alleen psychische klachten [The Outcome Questionnaire (OQ-45): A measure for psychiatric symptoms and more]. *De Psycholoog*, *40*, 53–63.

de Beurs, E., den Hollander-Gijsman, M. E., Helmich, S., & Zitman, F. G. (2007). The tripartite model for assessing symptoms of anxiety and depression: Psychometrics of the Dutch version of the mood and anxiety symptoms questionnaire. *Behaviour Research and Therapy*, *45*, 1609–1617.

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., … Zitman, F. G. (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 & Psychotherapy*, *18*, 1–12.

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.

de Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy*, *14*, 288–301.

Delgadillo, J., Overend, K., Lucock, M., Groom, M., Kirby, N., McMillan, D., … de Jong, K. (2017). Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy*, *99*, 89–97.

Derogatis, L. R. (1975a). *The brief symptom inventory*. Baltimore, MD: Clinical Psychometric Research.

Derogatis, L. R. (1975b). *The Symptom Checklist-90*. Baltimore, MD: Clinical Psychometric Research.

Dingemans, A. E., & van Furth, E. F. (2017). Het meten van verandering tijdens behandeling voor eetstoornissen: Een vergelijking van twee algemene en specifieke vragenlijsten [Measuring change during the treatment of eating disorders: A comparison of two types of questionnaires]. *Tijdschrift voor Psychiatry*, *59*, 278–285.

Fairburn, C. G., & Beglin, S. J. (2008). Eating Disorder Examination Questionnaire (EDE-Q 6.0). In C. G. Fairburn (Ed.), *Cognitive therapy and eating disorders* (pp. 309–313). New York, NY: Guilford Press.

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are'good'depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time… Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*, 1354–1366.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, *23*, 56–62.

Hansson, M., Chotai, J., Nordstöm, A., & Bodlund, O. (2009). Comparison of two self-rating scales to detect depression: HADS and PHQ-9. *British Journal of General Practice*, *59*, e283–e288.

Helmreich, I., Wagner, S., Mergl, R., Allgaier, A.-K., Hautzinger, M., Henkel, V., … Tadić, A. (2011). The Inventory of Depressive Symptomatology (IDS-C28) is more sensitive to changes in depressive symptomatology than the Hamilton Depression Rating Scale (HAMD17) in patients with mild major, minor or subsyndromal depression. *European Archives of Psychiatry and Clinical Neuroscience*, *261*, 357–367.

Husted, J. A., Cook, R. J., Farewell, V. T., & Gladman, D. D. (2000). Methods for assessing responsiveness. *Journal of Clinical Epidemiology*, *53*, 459–468.

Koeter, M. W., Ormel, J., & van den Brink, W. (1988). Totaalscore op de SCL-90 als maat voor de ernst van psychopathologie [SCL-90 total score as an index of severity of psychopathology]. *Nederlands Tijdschrift voor de Psychologie*, *43*, 381–387.

Kounali, D. Z., Button, K. S., Lewis, G., & Ades, A. E. (2016). The relative responsiveness of test instruments can be estimated using a meta-analytic approach: An illustration with treatments for depression. *Journal of Clinical Epidemiology*, *77*, 68–77.

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology*, *2*, 111–133.

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

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Volume 3: Instruments for adults* (3rd ed., pp. 191–234). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Mauriño, J., Cordero, L., & Ballesteros, J. (2012). The subjective well-being under neuroleptic scale–short version (SWN-K) and the SF-36 health survey as quality of life measures in patients with schizophrenia. *Patient preference and adherence*, *6*, 83–85.

McCrindle, B. W., Zak, V., Pemberton, V. L., Lambert, L. M., Vetter, V. L., Lai, W. W., … Cook, A. (2014). Functional health status in children and adolescents after Fontan: Comparison of generic and disease-specific assessments. *Cardiology in the Young*, *24*, 469–477.

Meesters, Y., Duijzer, W., Nolen, W., Schoevers, R., & Ruhé, H. (2016). Inventory of Depressive Symptomatology en verkorte versie in routine outcome monitoring van Stichting Benchmark (SBG). *Tijdschrift voor Psychiatrie*, *58*, 48–54.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, *134*, 382–389.

Reine, G., Simeoni, M.-C., Auquier, P., Loundou, A., Aghababian, V., & Lancon, C. (2005). Assessing health-related quality of life in patients suffering from schizophrenia: A comparison of instruments. *European Psychiatry*, *20*, 510–519.

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The inventory of depressive symptomatology (IDS): Psychometric properties. *Psychological Medicine*, *26*, 477–486.

Rush, J. A., Giles, D. E., Schlesser, M. A., Fulton, C. L., Weissenburger, J., & Burns, C. (1986). The inventory for depressive symptomatology (IDS): Preliminary findings. *Psychiatry Research*, *18*, 65–87.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., … Dunbar, G. C. (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.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., … de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*, 34–42.

Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research*, *12*, 349–362.

Tu, X. J., Hwang, W. J., Hsu, S. P., & Ma, H. I. (2017). Responsiveness of the short-form health survey and the Parkinson's disease questionnaire in patients with Parkinson's disease. *Health Qual Life Outcomes*, *15*, 75.

van der Does, A. (2002). *Handleiding bij de Nederlandse versie van Beck Depression Inventory—Second edition (BDI-II-NL) [manual for the BDI-II]*. Amsterdam, the Netherlands: Pearson.

van der Mheen, M., ter Mors, L. M., van den Hout, M. A., & Cath, D. C. (2018). Routine outcome monitoring bij de behandeling van angststoornissen: Diagnosespecifieke versus generieke meetinstrumenten. *Tijdschrift voor Psychiatrie*, *60*(1), 11–19.

Van, H. L., Dekker, J., Peen, J., van Aalst, G., & Schoevers, R. A. (2008). Identifying patients at risk of complete nonresponse in the outpatient treatment of depression. *Psychotherapy and Psychosomatics*, *77*, 358–364.

van Vliet, I. M., & de Beurs, E. (2007). Het Mini Internationaal Neuropsychiatrisch Interview (MINI) Een kort gestructureerd diagnostisch psychiatrisch interview voor DSM-IV- en ICD-10-stoornissen [The MINI-International Neuropsychiatric Interview. A brief structured diagnostic psychiatric interview for DSM-IV en ICD-10 psychiatric disorders]. *Tijdschrift voor Psychiatrie*, *49*, 393–397.

Wahl, I., Löwe, B., Bjorner, J. B., Fischer, F., Langs, G., Voderholzer, U., … Rose, M. (2014). Standardization of depression measurement: A common metric was developed for 11 self-report depression measures. *Journal of Clinical Epidemiology*, *67*, 73–86.

Walton, K., Ormel, J., & Krueger, R. (2011). *The dimensional nature of externalizing behaviors in adolescence: Evidence from a direct comparison of categorical, dimensional, and hybrid models*. Paper presented at the Journal of Abnormal Child Psychology. https://doi.org/10.1007/s10802-010-9478-y

Wardenaar, K. J., van Veen, T., Giltay, E. J., den Hollander-Gijsman, M., Penninx, B. W., & Zitman, F. G. (2010). The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *Journal of Affective Disorders*, *125*, 146–154.

Watson, D., Weber, K., Assenheimer, J. S., Clark, L. A., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology*, *104*, 3–14.

Wiebe, S., Guyatt, G., Weaver, B., Matijevic, S., & Sidwell, C. (2003). Comparative responsiveness of generic and specific quality-of-life instruments. *Journal of Clinical Epidemiology*, *56*, 52–60.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: de Beurs E, Vissers E, Schoevers R, Carlier IVE, van Hemert AM, Meesters Y. Comparative responsiveness of generic versus disorder-specific instruments for depression: An assessment in three longitudinal data sets. *Depress Anxiety*. 2019;36:93–102. https://doi.org/10.1002/da.22809