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

Edwin de Beurs¹ | Ellen Vissers² | Robert Schoevers² | Ingrid V. E. Carlier³ | Albert M. van Hemert³ | Ybe Meesters²

¹Faculty of Clinical Psychology, Leiden University, Leiden, The Netherlands
²Department of Psychiatry, University Medical Center Groningen, Groningen, The Netherlands
³Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands

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-head, 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, Nordstrom, & 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 (\( \text{ES}_{\text{EDE-Q}} = 0.84 \)) than on the BSI (\( \text{ES}_{\text{BSI}} = 0.45 \)). 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 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-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-SR 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 ($nth$) 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 $nth$ 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,
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 nth 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 ($\eta^2$) 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; $\eta^2$ 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 nth assessment). The correlations are of medium or high size (r > 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$_{BDI-TOT}$ = 0.96 and ES$_{BDI-COG}$ = 0.91); the SCL-TOT is less responsive (ES$_{SCL-TOT}$ = 0.75); and the SCL-DEP is somewhat more responsive (ES$_{SCL-DEP}$ = 0.82). Other subscales of the SCL-90 are again less responsive (ES$_{SCL-ANG}$ = 0.62 and ES$_{SCL-HOS}$ = 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

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

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

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Time</th>
<th>Instrument</th>
<th>Time x Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>df</td>
<td>F</td>
</tr>
<tr>
<td>BDI vs. SCL-TOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First, second, and third</td>
<td>160</td>
<td>2.318</td>
<td>111.64</td>
</tr>
<tr>
<td>BDI vs. SCL-DEP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First, second, and third</td>
<td>160</td>
<td>2.318</td>
<td>93.69</td>
</tr>
<tr>
<td>BDI vs. SCL-TOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vs. nth</td>
<td>233</td>
<td>1.232</td>
<td>181.50</td>
</tr>
<tr>
<td>BDI vs. SCL-DEP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vs. nth</td>
<td>233</td>
<td>1.232</td>
<td>178.77</td>
</tr>
</tbody>
</table>

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 η² in bold typeface.
ANOVA results 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 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 coefficients 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 ($ES_{OQ-SD} = 0.55$), followed by the OQ-TOT, the IDS-TOT, and the OQ-SR ($ES_{OQ-SR} = 0.48$). The OQ-IR is less responsive.

A subset of these data was analyzed with repeated measures ANOVAs. 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 $ES_{OQ-IR} = 0.42$.

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 ($ES_{MASQ-PA} = 1.02$); the BDI-TOT, BDI-SOM, BDI-AFF, MASQ-TOT, and the BSI-DEP ($ES_{BSI-DEP} = 0.79$) are all still fairly responsive. The BSI subscales not assessing depression are the least responsive ($ES_{BSI-SOM} = 0.40$).

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 $\eta^2$), which is illustrated in Figure 1 (right): the MASQ-PA and the BDI-TOT scale were the most responsive ($ES_{MASQ-PA} = 1.02$ and $ES_{BDI-TOT} = 0.92$); the BSI-DEP and the BSI-TOT were less responsive ($ES_{BSI-DEP} = 0.79$ and $ES_{BSI-TOT} = 0.71$). Almost all time and instrument contrasts are statistically significant. ES ($\eta^2$) 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).
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 change. 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

---

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

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Time Instrument</th>
<th>Time Instrument</th>
<th>Time x instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>df</td>
<td>F</td>
</tr>
<tr>
<td>IDS vs. OQ-TOT</td>
<td>First, second, third, and fourth</td>
<td>210</td>
<td>3,627</td>
</tr>
<tr>
<td>IDS vs. OQ-SD</td>
<td>First, second, third, and fourth</td>
<td>210</td>
<td>3,627</td>
</tr>
<tr>
<td>IDS vs. OQ-TOT</td>
<td>First vs. nth</td>
<td>832</td>
<td>1,831</td>
</tr>
<tr>
<td>IDS vs. OQ-SD</td>
<td>First vs. nth</td>
<td>832</td>
<td>1,831</td>
</tr>
</tbody>
</table>

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 η² in bold typeface.

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

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Time Instrument</th>
<th>Time Instrument</th>
<th>Time x instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>df</td>
<td>F</td>
</tr>
<tr>
<td>TOT scales</td>
<td>First to Fifth</td>
<td>479</td>
<td>4,1912</td>
</tr>
<tr>
<td>BDI, PA, BSI-DEP</td>
<td>First to Fifth</td>
<td>479</td>
<td>4,1912</td>
</tr>
<tr>
<td>TOT scales</td>
<td>First vs. nth</td>
<td>3,409</td>
<td>1,3408</td>
</tr>
<tr>
<td>BDI, PA, BSI-DEP</td>
<td>First vs. nth</td>
<td>3,409</td>
<td>2,3408</td>
</tr>
</tbody>
</table>

Notes. TOT-scales, total score on BDI, MASQ, and BSI; BDI, BDI-total score; PA, MASQ-positive affect; BSI-DEP, BSI-depression.
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-5 (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-Asberg 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 (sub)scales 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


SUPPORTING INFORMATION

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