Administration and Policy in Mental Health and Mental Health Services Research
https://doi.org/10.1007/s10488-024-01406-x

ORIGINAL ARTICLE

Designing a Measurement Feedback System for Personality Disorders: Should Outcome Monitoring be Based on Symptom Severity or Personality Functioning?

Marieke van Geffen[1,2] · Hester V. Eeren[3] · Joost Hutsebaut[1] · Odette Brand-de Wilde[4]

Accepted: 11 May 2024

Abstract
Measurement feedback systems (MFS) providing insight in treatment progress can improve mental healthcare outcomes. However, there is no uniform measurement feedback system that could be used to measure treatment progress for personality disorders (PD). This study compared two types of measures: a generic measure for symptom severity (Brief Symptom Index, BSI) and a specific measure for personality functioning (Severity Indices of Personality Problems, SIPP) at different points in time in order to provide insight in the most suitable measuring moment for a MFS for PD. This study is conducted in a sample of 996 Dutch PD patients (mean age 33.51 (SD 10.42), 73.1% female). Symptom severity and personality functioning were assessed before and multiple times during treatment, using a timespan of 24 months. Outcomes were examined over time using multilevel modeling. Symptom severity (generic measure) and personality functioning (specific measure) improved equally after 24 months. However, during these 24 months, different patterns of change were observed for symptom severity compared to severity of personality problems. In general, symptom severity decreased most during the 1st months of treatment, whereas personality functioning improved only after 6 months of treatment. A generic instrument of symptom severity is able to measure early changes in symptom distress but may not be able to measure longer term changes in personality functioning. The authors discuss policy implications for benchmarking using specific measures in the treatment of personality disorders.

Keywords Feedback systems · Benchmarking · Personality disorders · Symptom severity · Personality functioning

Introduction

Measurement feedback systems (MFS) have been introduced in health care organizations to monitor and improve the quality of care. They typically involve a systematic evaluation of the quality of care provided through a measurement system—usually based on Routine Outcome Monitoring (ROM) data—that is applied during treatment and that provides information for professionals and for the organization (Berwick et al., 2008; Bickman, 2008; Casparie et al., 1997; Harteloh, 2003; Hermann et al., 2000; Plas et al., 2001; Sluijs et al., 2007; Walburg & Brinkmann, 2001; Wollersheim et al., 2011). Previous studies have shown their promise in improving the quality of care for the individual patient (Lambert et al., 2005).

ROM data typically consist of systematically and repeatedly collected data on patients' mental health and functioning as an indicator of treatment outcome (Buwalda et al., 2011). The primary aim of ROM is usually to assess a patient's progress in treatment as a part of his or her treatment review (Lambert, 2007; Lambert et al., 2005). Consequently, professionals may change their treatment plan or policy, based upon the provided information. However, ROM data can also be used to provide transparency regarding the effectiveness of treatments in general as they may

✉ Marieke van Geffen
marieke.van.geffen@deviersprong.nl

1 De Viersprong, Beeklaan 2, 4661 EP Halsteren, The Netherlands

2 Department of Management Studies, Radboud University Nijmegen, Postbus 9108, 6500 HK Nijmegen, The Netherlands

3 Department of Psychiatry, Section Medical Psychology and Psychotherapy, Erasmus MC, PO box 2040, 3000 CA Rotterdam, The Netherlands

4 KieN, Van Kleffenslaan 1, 8442 CW Heerenveen, The Netherlands

Published online: 06 August 2024

allow a comparison of treatments within a mental healthcare institution or even between institutions. These comparative performance data can then be used to inform all stakeholders, such as mental health care organizations, insurance companies, and patients, about the quality of care of different providers (Barendregt, 2015; Buwalda et al., 2011). This so-called 'benchmarking' is defined as 'the continuous process of measuring products, services and practices against leaders, allowing the identification of best practices that will lead to measurable improvements in performance' (Bayney, 2005; Camp, 1989).

Although using MFS's can improve mental healthcare outcomes, these systems are not widely applied within mental health care (Rose & Bezjak, 2009). There are several barriers in applying these systems in clinical practice, such as resistance from stakeholders, organizations and professionals possibly resulting from a fear that information might reveal that treatment is ineffective (Bickman, 2008) and practical issues such as information needs to be distributed in an timely, efficient and uniform manner (Buwalda et al., 2011). Another important barrier is that the available measurements in mental healthcare are less precise compared to, for instance biomarkers in other areas of healthcare. Besides that, measurements in mental healthcare are also time consuming (Rose & Bezjak, 2009). Even more, the availability of many outcome measures of different domains or symptoms makes it hard to choose.

When using feedback systems, considerations for a suitable measure depends on the aim of the MFS, the burden of filling in such a measure by patients and the ease of use for professionals (Buwalda et al., 2011). A MFS can serve different aims such as assessing patient's progress in treatment or benchmarking treatment outcomes. These aims can conflict with each other as well as with the requirements each aim places on the measurement or available information (van Os et al., 2012, 2017). On the one hand, benchmarking benefits from identifying a generic outcome and timeframe, such as the percentage of patients with post-operation infections after the first hours of surgery. Stakeholders may compare different treatments or providers by comparing effect sizes based upon such a generic outcome. In the field of mental healthcare, (the reduction of) symptom severity is often used as an indication of treatment outcome for monitoring individual treatment progress and for benchmarking (de Beurs et al., 2015a, 2015b). On the other hand, assessing patients' progress in treatment may require a more specific outcome. For instance, most treatments for personality disorders (PDs) do not necessarily claim to primarily reduce symptom severity. These treatments aim to improve personality functioning (e.g., Bateman & Fonagy, 2004). According to recent theories, the core of PDs lies in impairments in self and interpersonal functioning on the one hand and in maladaptive personality traits on the other hand (Bender et al., 2011). While symptoms of personality and other disorders, along with the subjective burden they bring along, are assumed to be fluctuating, these underlying impairments and traits are thought to capture more durable aspects of a patient's psychopathology. They may account for the long-term consequences of PDs. Indeed, most specific treatment programs for PDs (i.e. Mentalization Based Treatment, Bateman & Fonagy, 2004, Transference-Focused Psychotherapy, Yeomans et al., 2002 and Schema therapy, Young, 1990; Young et al., 2003) target these personality impairments, rather than the presenting symptoms. They assume that changes in personality functioning may result in enduring changes in different areas of life. If so, such changes in personality functioning may be more reflective of successful treatment than a temporary relief of symptom stress. At the same time, this also raises the question whether a generic symptom severity index always provides the most valid base to assess individual treatment progress as well as to compare the quality of treatment for PDs between different institutions.

Besides the conflict of a generic versus a more specific outcome indicator, benchmarking and assessing individual treatment progress may also conflict regarding what a suitable measurement frequency would be. Benchmarking requires a generic timeframe. In Dutch Mental healthcare, the measurement frequency was set at a maximum of 1 year, starting at the moment a patient registers at a mental healthcare institution. However, assessing treatment progress might demand a longer time span. Improvements in personality functioning seem to require a longer period to become apparent (Perry et al., 1993) and might demand lengthy treatments or at least longer follow-up periods after treatment.

To investigate the possible misfit between a maximum time span of a year to measure outcome in the more lengthy treatment for PD patients and to investigate the question whether in PD patients a generic measure provides a good reflection of treatment response, the present study compared two types of outcome measures in a large sample of PD patients: a generic measure for symptom severity, the Brief Symptom Inventory (BSI; Derogatis, 1975) and a more specific measure of personality functioning, the Severity Indices of Personality Problems (SIPP, Verheul et al., 2008). Our first aim was to investigate both measures' ability to sensitively detect changes during PD-oriented treatment, hypothesizing that both indicators would show clear improvements making them suitable for MFS/ROM. Our second, and major aim was to compare the patterns of change for both measures throughout time. We hypothesized that both measures would show a different pattern of change over time. More specifically, we expected that a generic measure of symptom severity would reveal change earlier in time than a specific measure of personality functioning. Based upon the assumption

that personality changes require time, we expected changes in personality functioning to show a slower but longer trajectory of change. Finally, we explored whether these patterns of change may differ between different PD clusters. This may add to previous findings that patients with different types of personality disorders might show different patterns of change over time (Feenstra et al., 2014). Taken together, results from this study may inform a discussion on how to design an MFS for personality disorders.

Method

Participants and Procedure

Participants

ROM data collected at the Viersprong, institute for personality disorders and behavioral problems in the Netherlands, was used in this study.

From January 2012 up to December 2014, 3240 adult patients were referred for treatment at the Viersprong (Fig. 1). The Viersprong offers outpatient, day hospital and inpatient psychotherapy for patients with personality disorders cluster B, C and other specified (NOS). Of these patients, 75% (N = 2435) underwent a standard assessment as part of the intake procedure, including the Structured Clinical Interview for DSM-IV Axis I disorders (SCID-I) (First et al., 1997; translated by van Groenestijn et al., 1999) and the Structured Clinical Interview for DSM-IV Axis II Personality disorders (SCID-II) (First et al., 1996; translated by Weertman et al., 2000). These patients also completed several questionnaires as part of ROM. The data obtained from this initial assessment served as baseline data for our study. Based upon DSM-IV classification and therapists' expertise, patients were assigned to a relevant treatment program.

In total, 30% (N = 727) of the diagnosed patients did not meet criteria of a personality disorder and these patients were therefore excluded from the study sample. Of the remaining 1708 patients, 1686 patients (99%) were assigned to a treatment program within the Viersprong. Of these 1686 patients (40%) could not start treatment immediately and therefore were assigned to a waiting list. If their assigned treatment could not start within the selected timeframe (January 2012–December 2014), these patients were excluded from the study sample. However, treatment did not need to be finished in the selected timeframe. Also, treatment dropouts were not excluded. In total, 1012 patients were included in the study sample. 16 patients refused to fill in a questionnaire (2%), so 98% completed at least one questionnaire (BSI or SIPP) during the intake and treatment procedure. The final study sample thus consisted of 996 patients.

Of these 996 patients in the final study sample, 436 patients were diagnosed with a cluster B personality disorder as primary diagnoses, 206 patients had a primary diagnosis of a cluster C personality disorder and 354 patients had a personality disorder NOS (according to the SCID-II, corresponding with a 'Other specified PD' in DSM-5). No patients were diagnosed with a primary cluster A personality disorder (see Fig. 1). This distribution corresponds with the types of treatment being offered at the Viersprong. Patient could be offered different types of psychotherapy; mentalization based treatment (MBT), schema therapy (ST) or psychodynamic therapy (PDT). MBT was offered as an outpatient treatment, whereas ST and PDT were offered in outpatient and inpatient formats. Retrospective research on patients archived files does not require informed consent under Dutch Law. Because the questionnaires used in this study were part of the standard screening and treatment procedure and part

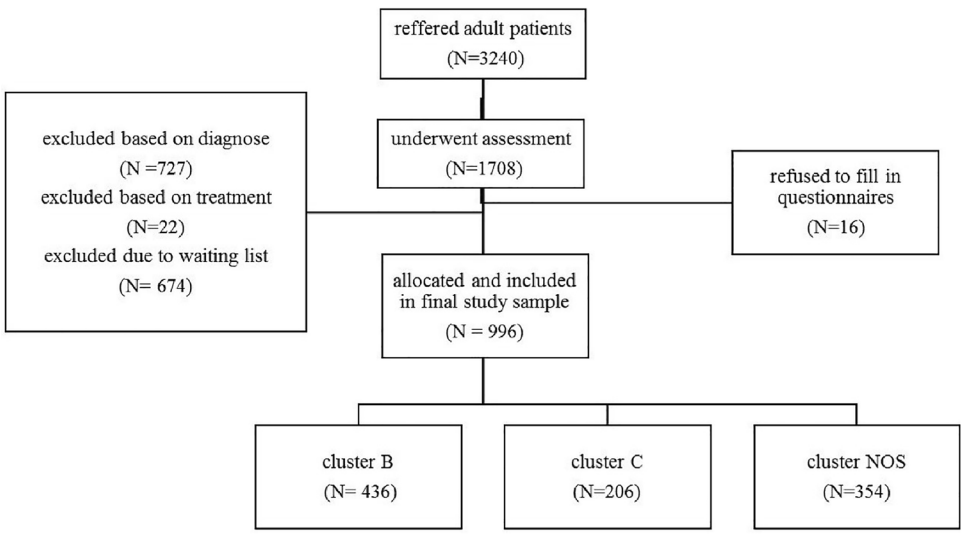

Fig. 1 Patient flow

of ROM, informed consent was not required. There were no payments made to the participants.

Procedure

All treatment programs at the Viersprong contain a standardized evaluation cycle of ROM. As part of this evaluation cycle, and as part of the ROM procedure, patients were asked to fill in self-report questionnaires. The number of times a patient was asked to complete these measures differed between every 6 weeks up to every 6 months (depending on the specific type of treatment). This was not necessarily related to the diagnosis of a patient, but to treatment setting. If a therapy was short and intensive, patients were asked to fill in these measures more frequently. However, as all treatments aimed to reduce the severity of psychological symptoms (BSI) and to improve personality functioning (SIPP), both measures were included in all sets.

All completed questionnaires of the 996 patients in the study sample—completed between January 2012 and July 2015—were used in the statistical analyses. Duplicate questionnaires were removed. Because treatment evaluation was not repeated within a period shorter than 6 weeks, a questionnaire was considered as a duplicate when this questionnaire was completed twice a month.

Only outcome data of the first 24 months were used, as most of the treatments finished within 24 months. Of the 996 patients in the study sample, a total of 3430 BSI and 3346 SIPP questionnaires were completed and included within the time frame of 24 months. Figure 2 shows the number of questionnaires that were completed per month. All data was collected during treatment. No follow-up data was included.

Instruments

Outcome Measures

The severity of psychological symptoms was measured using the Dutch version of the Brief Symptom Inventory (BSI, Derogatis, 1975). The BSI is a self-report scale consisting of 53 items describing general symptoms of psychopathology. The BSI is a multidimensional instrument covering nine primary symptom clusters (somatization, obsession-compulsion, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism) and it yields three global indices of severity: Positive Symptom Distress Index, Positive Symptom Total, and Global Severity Index (GSI). GSI scores range from 0 to 4, with higher scores indicating a higher level of psychological and emotional distress (de Beurs, 2011; Derogatis, 1975). The BSI-items refer to severity of psychological symptoms during the past week.

Personality functioning was measured using the Severity Indices of Personality Problems, which is originally a Dutch questionnaire (Verheul et al., 2008). The SIPP is a dimensional self-report measure and aims to measure the severity of the generic and changeable components of personality disorders. The response categories range from 1 to 4 and are described as "fully disagree," "partly disagree," "partly agree," or "fully agree." The measure comprises five higher-order domains that relate to impairments in self and interpersonal functioning; Self-Control (Slfc); the capacity to tolerate, use and control emotions and impulses, Identity Integration (II); the capacity to consider oneself and one's own life as stable, integrated and meaningful, Responsibility (Resp); the capacity to achieve realistic goals, Relational

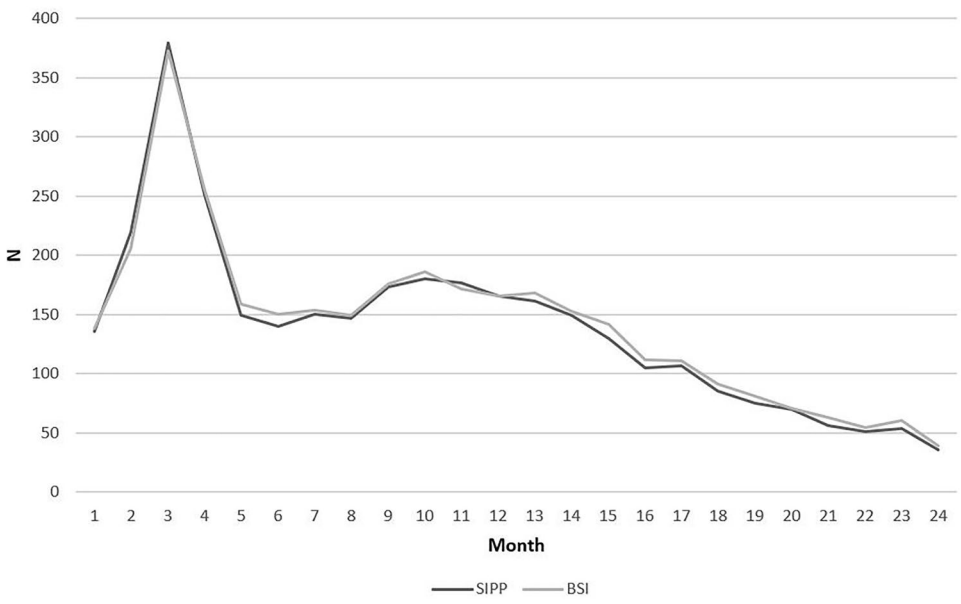

Fig. 2 Number of BSI and SIPP questionnaires per month

Capacities (Rel); the capacity to love others and feel loved; and Social Concordance (Soc); the capacity to appreciate the boundaries of others, to control aggressive impulses and to cooperate with others. The SIPP does not generate an overall score of personality functioning. High scores on the SIPP indicate more adaptive functioning. In the ROM data used in this study, two versions of the SIPP were used: a short form of 60 items and a long form of 118 items. The SIPP short form can be extracted from the long form to create the short form (Arnevik et al., 2009). After extracting the SIPP-sf from the SIPP-118, the domains of the SIPP-60 were calculated..

Statistical Analyses

The BSI and SIPP outcomes were examined over time using multilevel modeling, where time was modelled in months after the start of the treatment (i.e. Feenstra et al., 2014). We used multilevel modeling to deal with the dependency of repeated measurements on the same subject in time, and to deal with unbalanced time points and missing observations at certain time points (Hox, 2002). Treatment effects were estimated at several time points during and after treatment, depending on the treatment (see Fig. 2). We first postulated a saturated model with intercept and slope as random variables and with time as Level I and patients as Level II. To decide on the covariance structure and to assess whether the slope could be added as a random slope, we used the deviance statistic using restricted maximum likelihood (Hox, 2002; Peugh, 2010; Singer & Willett, 2003). In the fixed part of the model, time, time squared, a cubic term of time and the logarithm of time were entered as independent variables. Entering time as time squared, cubic time and logarithmic time allows for different types of trends of time in the model. Non-significant fixed time effects were excluded in an iterative process until a parsimonious model was reached that did not significantly differ from the saturated model. To decide on statistical significance, we used the deviance statistic using ordinary maximum likelihood (Hox, 2002; Peugh, 2010; Singer & Willett, 2003). Cohen's D effect sizes with pooled standard deviations from the models were used to describe the changes from baseline to 24 months after the start of treatment (Cohen, 1992). An effect size of 0.2 is considered to be small, a medium effect size is defined as 0.5, while 0.8 is considered to be a large effect size (Cohen, 1992). We investigated the effect size of six outcomes: the total score of symptom severity (BSI), and five scales of the SIPP. We explored if trajectories of change differed between different types of diagnoses. We added dummy variables for each cluster to the fixed effects of the model, which were defined as being or not being diagnosed with each of these personality disorders clusters. We also added their interaction with time, time squared, a cubic term of time and the logarithm of time to the saturated model. We followed the steps in modelling the outcomes over time in the multilevel models as described. When removing non-significant effects, the interaction terms were treated as nested under their main effects (Hox, 2002). All analysis were conducted in SPSS.

Results

Sample Characteristics

The mean age of the patients included in the study sample was 33.51 (SD 10.42). In total, 52.6% of the patients were referred by first or second line treatment centers (N = 524), while 33.8% was referred to the Viersprong by specialized care centers (N = 337). For 13.6% of the patients the referral agency was unknown (N = 135). Baseline characteristics of the study sample are presented in Table 1.

Treatment Outcomes Over Time

The first aim of the study was to study trajectories of change for different types of outcome measures (generic versus specific) for personality disorders (generic versus specific) in terms of their effect sizes during the first 24 months of treatment. Patients showed significant improvements at the end of the 24 months of treatment compared to their baseline scores on all six, generic and specific, outcome measures. Patients reported significantly less symptom severity and significantly less personality problems after 24 months of treatment. The total effect sizes ranged from 0.40 to 0.88, indicating medium to large effects (Cohen, 1988) (see Table 2, defined in the column 'd total').

Further, we compared the patterns of change for both indicators—psychological symptom severity reduction and reduction of personality problems during treatment—over time. If the effect size is estimated per 6 months of time

Table 1 Demographic variables of the study sample (N = 996)

Baseline characteristics	Study sample (N = 996)
Age M (SD)	33.51 (10.42)
Female N (%)	73.1%
Referred by	
General practice organizations/agencies	52.7%
Specialized care organizations/agencies	33.7%
Unknown	13.6%
Year of referral	
2012	27.7%
2013	35.9%
2014	36.3%

Table 2 Effect sizes per indicator per 6 months

	Baseline M	T6* M (SD)	*d**	T12 M (SD)	*d*	T18 M (SD)	*d*	T24 M (SD)	*d*	*d total***
BSI	1.91	1.55 (.74)	− .50	1.44 (.72)	− .15	1.24 (.76)	− .26	1.16 (.86)	− .09	− .88
SIPP Slfc	2.39	2.62 (.74)	.31	2.76 (.70)	.19	2.97 (.70)	.31	3.02 (.74)	.06	.81
SIPP II	1.97	2.11 (.64)	.23	2.32 (.67)	.31	2.57 (.77)	.33	2.69 (.93)	.13	.87
SIPP Resp	2.67	2.82 (.63)	.23	2.87 (.61)	.08	3.01 (.61)	.23	2.99 (.64)	− .03	.48
SIPP Rel	2.25	2.26 (.63)	.03	2.38 (.64)	.19	2.59 (.69)	.30	2.60 (.77)	.01	.48
SIPP Soc	2.78	2.91 (.61)	.22	2.95 (.58)	.05	3.07 (.58)	.21	3.04 (.62)	− .04	.40

*T6 = estimated mean, standard deviation and effect size at 6 months compared to baseline, T12 = 12 months compared to 6 months, T18 = 18 months compared to 12 months, T24 = 24 months compared to 18 months

**d Total represents the effect size of 24 months compared to baseline

(thus, the effect size of the effect at 6 months compared to the baseline score, 12 months compared to 6 and 18 months compared to 12 months), Table 2 shows that there are indeed differences between the indicators. For instance, the effect size of the BSI at 6 months compared to baseline is − 0.50 (a decrease of symptom severity) while the effect sizes of the SIPP domains at 6 months only range between 0.03 and 0.31 (an increase of personality functioning, signaling more adaptive functioning). However, the effect size of the BSI at 12 months compared to 6 months is 0.15 while the corresponding effect sizes of the SIPP ranges between 0.05 and 0.31. All estimated effect sizes are presented in Table 2. Table 2 shows the effect sizes per 6 months, to give insight into the pattern of change over time.

We also explored changes in symptom severity and personality functioning over time when taking into account different clusters of PD diagnoses. All clusters (B, C and NOS) of personality disorders showed significant improvements after 24 months of treatment, compared to baseline, on all outcome indicators (Table 3, column 'd total'). The total effect sizes ranged from 0.48 to 1.07 for cluster B, indicating medium to large effects. For cluster C the effect sizes ranged from 0.12 to 0.74 and for Personality Disorder NOS from 0.09 to 0.72, indicating low to medium effects.

Table 3 Effect sizes per cluster of personality disorders

Cluster		Baseline M	T12* M (SD)	*d**	T24 M (SD)	*d*	*d total***
B	BSI	2.07	1.60 (.72)	− .62	1.26 (.83)	− .42	− .96
N = 436	SIPP Self-control	2.11	2.37 (.64)	.40	2.86 (.67)	.74	1.07
	SIPP Identity integration	1.87	2.10 (.67)	.36	2.63 (.84)	.62	.93
	SIPP Responsibility	2.59	2.70 (.61)	.18	2.90 (.62)	.32	.48
	SIPP Relational capacities	2.16	2.25 (.64)	.14	2.58 (.73)	.44	.57
	SIPP Social concordance	2.55	2.70 (.56)	.27	2.90 (.58)	.36	.59
C	BSI	1.51	1.28 (.72)	− .32	0.90 (.82)	− .47	− .74
N = 206	SIPP Self-control	2.95	3.19 (.64)	.38	3.45 (.67)	.39	.71
	SIPP Identity integration	2.10	2.27 (.67)	.26	2.52 (.84)	.30	.52
	SIPP Responsibility	2.77	2.99 (.61)	.35	3.17 (.62)	.30	.63
	SIPP Relational capacities	2.25	2.21 (.64)	− .07	2.33 (.73)	.18	0.12
	SIPP Social concordance	3.17	3.03 (.56)	− .26	2.94 (.58)	− .14	− .39
NOS	BSI	1.63	1.32 (.72)	− .42	1.09 (.82)	− .28	− .64
N= 354	SIPP Self-control	2.68	2.93 (.64)	.38	3.00 (.67)	.12	.46
	SIPP Identity integration	2.18	2.45 (.67)	.40	2.77 (.84)	.38	.72
	SIPP Responsibility	2.80	2.98 (.61)	.30	3.00 (.62)	.04	.32
	SIPP Relational capacities	2.46	2.48 (.64)	.04	2.65 (.73)	.23	.26
	SIPP Social concordance	3.09	3.03 (.56)	− .10	3.14 (.58)	.19	.09

*T12 = estimated mean, standard deviation and effect size at 12 months compared to baseline, T24 = 24 months compared to 12 months

**d total represents the effect size of 24 months compared to baseline

When considering the effect sizes per year, thus 12 months compared to baseline (T12 in Table 3) and 24 months compared to 12 months (T24 in Table 3), differences between the indicators can be observed for patients with a cluster B personality disorder. Whereas symptom severity showed a big change in the 1st year (T12 in Table 3) with an effect size of − 0.62 compared to an effect size of − 0.42 after the 2nd year of treatment (T24 in Table 3). Indicators of personality functioning showed an opposite pattern of change, namely a larger change during the 2nd year (for instance Self Control cluster B effect size of 0.74, T24 in Table 3) compared to the 1st year (effect size of 0.40, T12 in Table 3).

To provide insight in the pattern of change of psychological symptom severity and personality problems severity per cluster of personality disorders during the selected 24 months, the effect sizes are visualized in Figs. 3, 4 and 5. In contrast to Tables 2 and 3, these figures show the effect sizes per month compared to baseline. Figure 3 shows the results of patients diagnosed with cluster B personality disorder, Fig. 4 for cluster C and Fig. 5 for NOS personality disorders. The figures show a different pattern of change for symptom severity compared to personality problems. Symptom severity decreases most during the first 6 months of treatment whereas personality problems show the biggest improvements between 6 and 18 months after starting treatment. However, these differences in patterns of change are most visible for patients with cluster B personality disorder (Fig. 3) as compared to patients with a cluster C (Fig. 4) or personality disorder NOS (Fig. 5). Patients with a cluster C and NOS personality disorder show the most improvement for symptom severity and personality problems during the first 19 months of treatment.

Discussion

Though measurement feedback systems can be used to monitor and improve quality of care, there is no agreement on the measure and timespan that should be used to measure treatment progress in a MFS for personality disorders. As different measurements and/or different timespans may give a different outcome of treatment effect, we aimed to compare a generic and specific outcome for measuring PD treatment outcome. Using repeated measures within a timeframe of 24 months, we found that both measures revealed significant improvements within this treatment period with medium to large effect sizes for both outcomes. After 24 months, only minor differences were observed between both types of indicators, suggesting that when using this timeframe, conclusions based on both measures may be similar. However, different trajectories of change were observed within these 24 months, with symptom severity showing the largest improvement within the first 6 months of treatment whereas personality functioning showing its major change between 6 and 18 months. This finding confirmed our hypothesis that symptom relief seem to be more readily achieved than improvements in personality functioning. Finally, our exploration of potential differences between the clusters of PDs revealed that this finding was the clearest for Cluster B PDs.

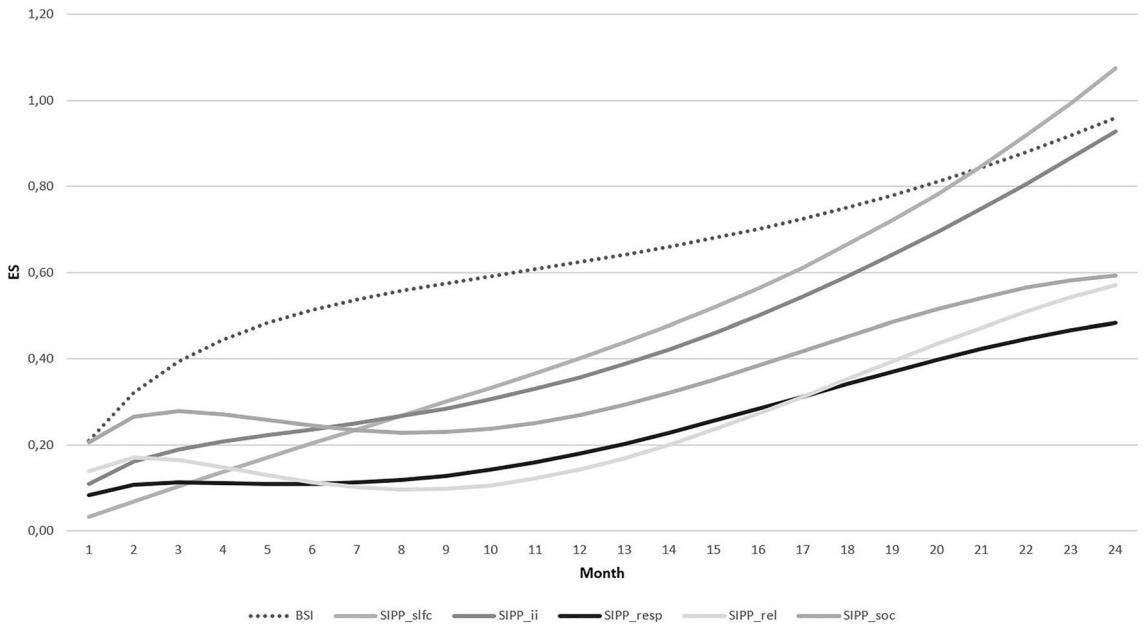

Fig. 3 Pattern of effect sizes for cluster B

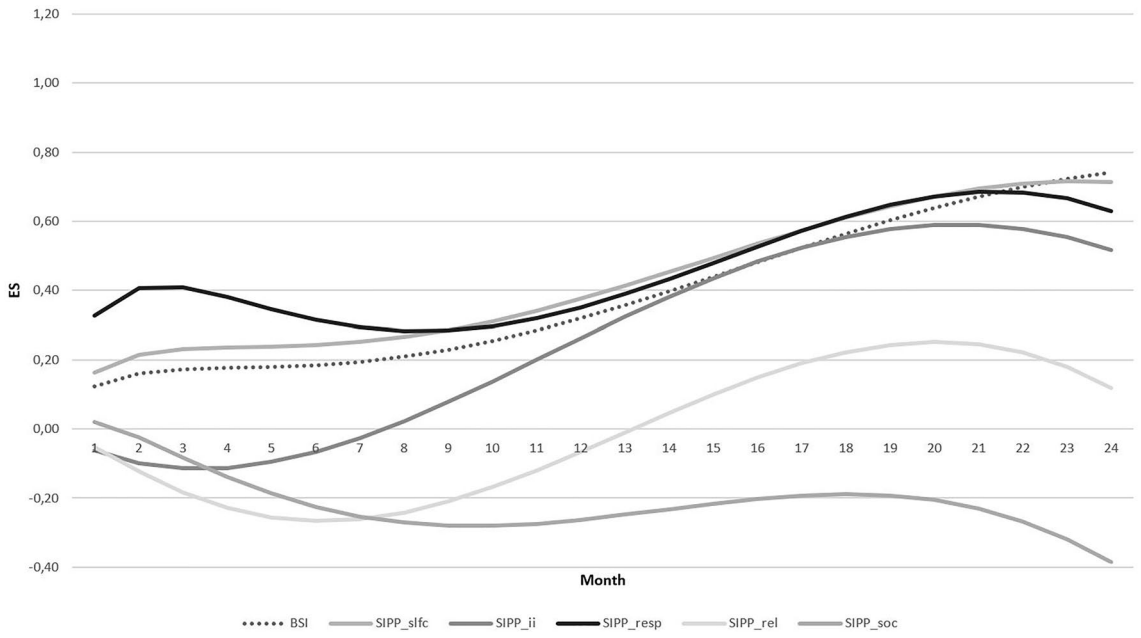

Fig. 4 Pattern of effect sizes for cluster C

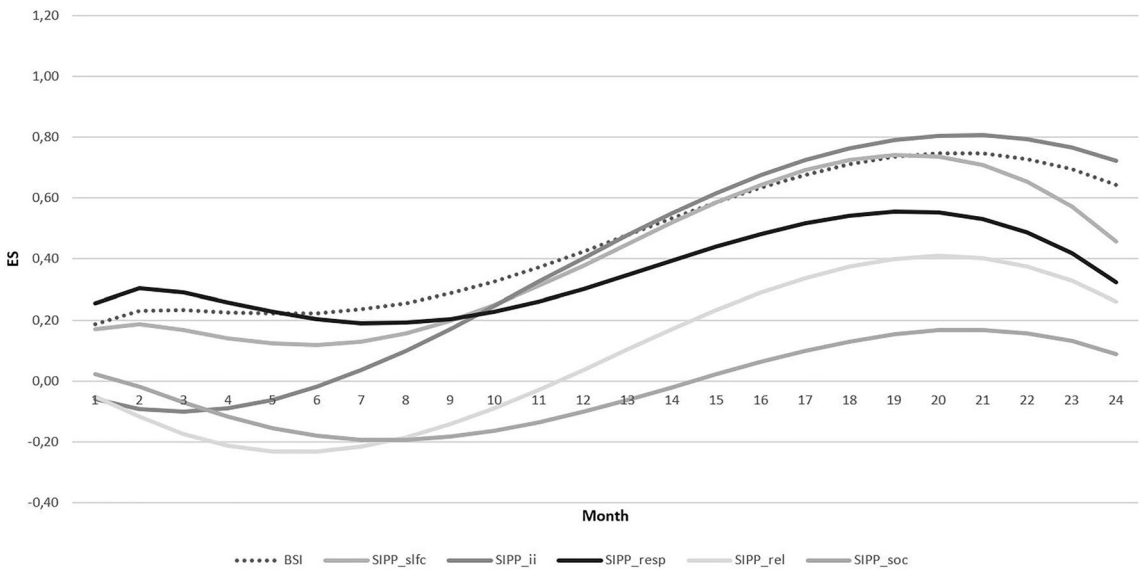

Fig. 5 Pattern of effect sizes for cluster NOS

Firstly, these findings may have important implications when selecting measures for a MFS aimed at monitoring individual outcomes. They suggest that using a generic symptom measure to estimate progress in treatment may present a too optimistic picture of a patient's progress after a short (i.e., 6 months) period of time. Though it can motivate patients if they observe progress quickly in treatment, it could also lead to finishing treatment too early. A consequence may be that relevant changes in personality functioning have not yet been achieved since our data show the biggest changes in PD functioning after 6 months of treatment, which may also affect the sustainability of symptom relief. Most evidence-based treatments for PDs require at least 12 to 18 months of treatment (Storebø et al., 2020), which is consistent with our finding that relevant changes in PD are observed later in treatment.

Secondly, our findings may have important consequences for a MFS aimed at benchmarking. If the chosen time frame

is long enough, i.e. 2 years, no different conclusions on treatment effect would be drawn when using a symptom or personality measure. Both measures indeed show highly comparable results after 2 years. Using a shorter time frame of 1 year—which has been the standard in the Dutch mental health care—may indeed lead to different conclusions when using a symptom or a personality measure. However, even if a timeframe of 2 years is used, it can be relevant to use both measures. For example, changes in personality functioning may be more relevant for predicting a sustainable treatment effect (possibly also in symptom severity) than symptom changes assessed by a general symptom severity index (Weekers, et al., 2024). Previous studies have indeed shown that relapses of mental state disorders, like depression, seem to be more likely when personality problems are not targeted in treatment (e.g. Oleski et al., 2012). Using a marker of change in personality functioning may therefore highlight different changes compared to changes in symptom severity only. A recommendation would be to add personality functioning measures in future studies on symptom disorders and symptom change over time. Finally, as different clusters of personality disorders showed different patterns of change, for benchmarking reasons it could be relevant to distinguish between these clusters.

Despite its relevance, some limitations of the underlying study need to be mentioned. Firstly, though we used data of a large number of patients, including a large number of outcomes over time, data was solely derived from patients at the Viersprong. No patients with cluster A PD were included in the sample, as no treatment was offered for them. Also, a high number of patients were diagnosed with personality disorder not otherwise specified (NOS, now called Other Specified PDs). It remains unclear whether patients from other institutes would show the same pattern of change. A replication study on a different data set could therefore be subject of future research. Secondly, only data of the first 24 months of treatment were included. Due to waiting list problems, most treatments did not start within the 1st month after assignment, therefore, some treatments were not finished after the included 24 months. Our study results therefore do not provide insights in overall treatment effects. Thirdly, the different timespans that's been used in both instruments could have impacted the results: the BSI asks respondents to report about the last week, while the SIPP asks about the last 3 months. Fourthly, the timeframes and inclusion criteria were based on common practice of benchmarking in the Netherlands. As the Dutch reimbursement is based on timeframes of 1 year and drop-outs are included in benchmarking, our design was based on these guidelines. This common practice can however create a potential self-selection bias as different samples of patients can be selected at different time intervals. Baseline differences are not controlled for in the analyses. This method is based on the common practice of benchmarking in the Netherlands. Benchmarking in the Netherlands is based on all clients who received treatment, making no distinction between clients who may only have received one consult or who received intensive inpatient treatment. Other countries might apply different benchmark strategies. Fifth, treatment settings were mixed, i.e. outpatient, day hospital an inpatient. The impact of these settings and different treatment models on the results was not studied. Therefore, we do not know if the results can be generalized to each of these settings. Finally, we want to stress that our focus was solely on the different trajectories of symptom and personality change patterns and their potential implications. As we don't have information on whether actual feedback was provided based upon these ROM data, neither on potential therapist effects regarding the provision of feedback, many clinically relevant issues that would focus more on the clinical applicability and value of ROM data and feedback for review of progress in treatment and altering the treatment plan, couldn't be investigated. Similarly, our focus was not on potential differences between treatment modalities or methods.

Taken together, our study provides insight into two severity indices that are commonly used to give information on treatment progress in feedback systems. It indicated that general symptom severity and personality functioning provide the same results after 24 months but show different patterns of change during this timeframe. This might be of particular interest for those who are concerned with designing an MFS for PD. This study shows different choices can be made regarding the measurement and timeframe depending on the aim of the MFS. This study might also be relevant for stakeholders, such as health insurance companies, since treatment for personality impairments is often more complex and lengthier, and therefore more expensive than treatment for mental state disorders (NICE, 2009). Also, if changes in personality functioning would be predictive of more durable improvement, stakeholders might be interested in this later treatment outcome, since its cost might pay off over a longer period of time. Future studies may therefore investigate the value of both measures in predicting sustainable recovery and whether additional treatment is prevented by treating personality functioning.

Author Contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by Marieke van Geffen and Hester van Eeren. Data analysis was done by Hester van Eeren. The first draft of the manuscript was written by Marieke van Geffen. Review, Editing and Supervision was done by Joost Hutsebaut and Odette Brand and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability The data that support the findings of this study are available from the corresponding author, M. van Geffen, upon request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Consent to Participate This distribution corresponds with the types of treatment being offered at the Viersprong. Retrospective research on patients archived files does not require informed consent under Dutch Law. Because the questionnaires used in this study were part of the standard screening and treatment procedure and part of ROM, informed consent was not required. There were no payments made to the participants.

References

Arnevik, E., Wilberg, T., Monsen, J. T., Andrea, H., & Karterud, S. (2009). A cross-national validity study of the severity indices of personality problems (SIPP-118). *Personality and Mental Health, 3*, 41–55.

Barendregt, M. (2015). Benchmarken en andere functies van ROM: Back to basics. *Tijdschrift Voor Psychiatrie, 57*(7), 517–525.

Bateman, A. W., & Fonagy, P. (2004). *Psychotherapy for borderline personality disorder: Mentalization-based treatment*. Oxford University Press.

Bayney, R. (2005). Benchmarking in mental health: An introduction for psychiatrists. *Advances in Psychiatric Treatment, 11*(4), 305–314.

Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in DSM-5, Part I: A review of theory and methods. *Journal of Personality Assessment, 93*, 333–346. https://doi.org/10.1080/00223891.2011.583808

Berwick, D. M., Nolan, T. W., & Whittington, J. (2008). The triple aim: Care, health, and cost. *Health Affairs, 27*(3), 759–769. https://doi.org/10.1377/hlthaff.27.3.759

Bickman, L. (2008). A measurement feedback system (MFS) is necessary to improve mental health outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*(10), 1114–1119. https://doi.org/10.1097/CHI.0b013e3181825af8.A

Buwalda, V. J. A., Nugter, M. A., Swinkels, J. A., & Mulder, C. L. (2011). Praktijkboek ROM in de ggz Een leidraad voor gebruik en implementatie van meetinstrumenten. Retrieved from https://www.tijdstroom.nl/boek/praktijkboek-rom-in-de-ggz#.VMDlBUeG9Ik

Camp, R. C. (1989). *Benchmarking: The search for industry best practice*. ASQC Press.

Casparie, A. F., Sluijs, E. M., Wagner, C., & de Bakker, D. H. (1997). Quality systems in Dutch health care institutions. *Health Policy (amsterdam, Netherlands), 42*, 255–267.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

de Beurs, E. (2011). *Brief symptom inventory-BSI-, Brief symptom inventory 18, -BSI 18-, Handleiding herziene editie 2011*. PITS B.V.

de Beurs, E., Barendregt, M., de Heer, A., van Duijn, E., Goeree, B., Kloos, M., Kooiman, K., Lionarons, H., & Merks, A. (2015a). Comparing methods to denote treatment outcome in clinical research and benchmarking mental health care. *Clinical Psychology & Psychotherapy*. https://doi.org/10.1002/cpp.1954

de Beurs, E., Barendregt, M., Rogmans, B., Robbers, S., van Geffen, M., van Aggelen-Gerrits, M., & Houben, H. (2015b). Denoting treatment outcome in child and adolescent psychiatry: A comparison of continuous and categorical outcomes. *European Child and Adolescent Psychiatry, 24*(5), 553–563. https://doi.org/10.1007/s00787-014-0609-9

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

Feenstra, D. J., Laurenssen, E. M. P., Timman, R., Verheul, R., Busschbach, J. J. V., & Hutsebaut, J. (2014). Long-term outcome of inpatient psychotherapy for adolescents (IPA) with personality pathology. *Journal of Personality Disorders, 28*(5), 637–656. https://doi.org/10.1521/pedi_2014_28_132

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured clinical interview for DSM-IVAxis I disorders (SCID I)*. American Psychiatric Press.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., & Benjamin, L. (1996). *Structured clinical interview for DSM-IV axis II personality disorders (SCID II)*. American Psychiatric Press.

Harteloh, P. P. M. (2003). Quality systems in health care: A sociotechnical approach. *Health Policy, 64*, 391–398.

Hermann, R. C., Leff, H. S., Palmer, R. H., Yang, D., Teller, T., Provost, S., Jakubiak, C., & Chan, J. (2000). Quality measures for mental health care: Results from a national inventory. *Medical Care Research and Review, 57*, 136–154. https://doi.org/10.1177/1077558700573008

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications* (1st ed.). Lawrence Erlbaum Associates.

Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*(1), 1–14. https://doi.org/10.1080/10503300601032506

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice sugges tions. *Journal of Clinical Psychology, 61*(2), 165–174. https://doi.org/10.1002/jclp.20113

National Institute for Health and Clinical Excellence. (2009). *Borderline personality disorder: Treatment and management*. British Psychological Society & The Royal College of Psychiatrists.

Oleski, J., Cox, B. J., Robinson, J., & Grant, B. (2012). The predictive validity of cluster C personality disorders on the persistence of major depression in the national epidemiologic urvey on alcohol and related conditions. *Journal of Personality Disorders, 26*, 322–333. https://doi.org/10.1521/pedi.2012.26.3.322

Perry, J. C. (1993). Longitudinal studies of personality disorders. *Journal of Personality Disorders, 1*, 63–85.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*, 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Plas, M., Hofhuis, H., & van den Ende, E. (2001). *Implementatie kwaliteitsbeleid paramedische zorg: Aansturing en organisatie van het IKPZ*. NIVEL.

Rose, M., & Bezjak, A. (2009). Logistics of collecting patientreported outcomes (PROs) in clinical practice: An overview and practical examples. *Quality of Life Research, 18*(1), 125–136.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Sluijs, E., Keijser, A., & Wagner, C. (2007). *Kwaliteitssystemen in zorginstellingen de stand van zaken in 2005*. NIVEL.

Storebø, O. J., Stoffers-Winterling, J. M., Völlm, B. A., Kongerslev, M. T., Mattivi, J. T., Jørgensen, M. S., Faltinsen, E., Todorovac, A., Sales, C. P., Callesen, H. E., Lieb, K., & Simonsen, E. (2020). Psychological therapies for people with borderline

personality disorder. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD012955

van Groenestijn, M. A. C., Akkerhuis, G. W., Kupka, R. W., Schneider, N., & Nolen, W. A. (1999). *Gestructureerd klinisch interview voor de vaststelling van DSM-IV as I stoornissen (Structured Clinical Interview for DSM-IV Axis I Disorders)*. Swets & Zeitlinger.

Van Os, J., Berkelaar, A., & Hafkenscheid, A. E. (2017). Benchmarken: Doodlopende weg onder het mom van 'ROM. *Tijdschrift Voor Psychiatrie, 59*, 247–250.

Van Os, J., Kahn, R., Denys, D., Schoevers, R. A., Beekman, A. T. F., Hoogendijk, W. J. G., Van Hemert, A. M., Hodiamont, P. P. G., Scheepers, F., Delespaul, P. A., & Leentjens, A. F. G. (2012). ROM: Gedragsnorm of dwangmaatregel? Overwegingen bij het themanummer over routine outcome monitoring. *Tijdschrift Voor Psychiatrie, 54*, 245–253.

Verheul, R., Andrea, H., Berghout, C. C., Dolan, C., Busschbach, J. J., van der Kroft, P. J., Bateman, A. W., & Fonagy, P. (2008). Severity Indices of Personality Problems (SIPP-118): Development, factor structure, reliability, and validity. *Psychological Assessment, 20*(1), 23–34. https://doi.org/10.1037/1040-3590.20.1.23

Walburg, J. A., & Brinkmann, J. (2001). *Kwaliteit is geen toeval: De praktijk van de kwaliteitszorg in de ggz*. GGZ Nederland.

Weekers, L. C., Hutsebaut, J., Rovers, J. M. C., & Kamphuis, J. H. (2024). Head-to-head comparison of the alternative model for personality disorders and Section II personality disorder model in terms of predicting patient outcomes 1 year later. *Personality Disorders, 15*(2), 101–109. https://doi.org/10.1037/per0000637

Weertman, A., Arntz, A. & Kerkhofs, M. L. M. (2000). *Gestructureerd diagnostisch interview voor DSM-IV persoonlijkheidsstoornissen (SCID-II)*. Swets Test Publishers.

Wollersheim, H., Bakker, P., & van der Weijden, T. (2011). Inleiding in kwaliteit en veiligheid van zorg 1. *Kwaliteit en veiligheid in patiëntenzorg* (pp. 13–24). Bohn Stafleu van Loghum.

Yeomans, F. E., Clarkin, J. F., & Kernberg, O. F. (2002). *A primer of transference focused psychotherapy for the borderline patient*. Jason Aronson Inc.

Young, J. E. (1990). *Cognitive therapy for personality disorders*. Professional Resources Press.

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2003). *Schema therapy: A practitioners guide*. Guildford Press.