Effects of Routine Outcome Monitoring (ROM) on therapy outcomes in the course of an implementation process. A randomized clinical trial

Heidi Brattland[1, 2], John Morten Koksvik [1, 2], Olav Burkeland [2], Rolf Wilhelm Gråwe [1], Christian Klöckner [3], Olav Morten Linaker [1], Truls Ryum [3], Bruce Wampold [4], Mariela Loreto Lara-Cabrera [2*], Valentina Cabral Iversen [1,2*]

[1] Department of Mental Health, Faculty of Medicine, The Norwegian University of Science and Technology (NTNU), Trondheim, Norway ; [2] Department of Tiller DPS, St Olavs University Hospital, Trondheim, Norway; [3] Department of Psychology, Faculty of Social Sciences and Technology Management, The Norwegian University of Science and Technology (NTNU), Trondheim, Norway; [4]Research Institute, Modum Bad Psychiatric Center, Vikersund, Norway and Department of Counseling Psychology, University of Wisconsin-Madison, Madison, WI, USA; [*]Shared last author credit

Acknowledgments. We are grateful to the Liaison Committee for Education, Research and Innovation in Central Norway and the Tiller Mental Health Center, St. Olavs Hospital for funding this research, to all participating clients and therapists, and to Scott D. Miller and Rolf Sundet for lending their support, experience and wisdom to this project.

Objective: This study investigated the effects of the Partners for Change Outcomes Management System (PCOMS) in adult out-patient treatment at a hospital-based mental health clinic. It also investigated whether the effects differed with the timing of the treatment within a four-year implementation period, with clients' initial distress levels, and between therapists. ***Method:*** Adult clients (N = 170) were randomized to treatment as usual (TAU) or Routine Outcome Monitoring (ROM). Twenty therapists provided therapy in both conditions. Therapy outcome was measured by the Behavior and Symptoms Identification Scale (BASIS-32). Data were analyzed in a series of multilevel models. ***Results:*** Clients in the ROM condition were 2.5 times more likely to demonstrate improvement than those in the TAU condition. Controlling for therapist variability, the overall effect size in favor of ROM was small (d = 0.26, p =.037). The superiority for ROM over TAU increased significantly over the duration of the study. ROM effects were not moderated by clients' initial distress levels. Differences between therapists accounted for 9%–10% of the variability in outcomes, and there were no significant differences in ROM effects between therapists. ***Conclusions:*** ROM was associated with better treatment outcomes independent of clients' initial distress levels. Clients treated later in the study benefitted more from ROM than those treated earlier.

What is the public health significance of this article? This study demonstrated improved outcomes for adult outpatient treatment when clients' session-to-session treatment responses were tracked with short Routine Outcome Monitoring (ROM) questionnaires. The effect of ROM increased over the four-year trial period, suggesting that it may take time to implement an effective ROM program and that training and supervision of therapists should be sustained over time.

Keywords: Feedback, implementation, psychotherapy, Routine Outcome Monitoring, therapist effects

Routine Outcome Monitoring, ROM[1] utilizes client self-report measures to monitor treatment responses throughout therapy and to alert therapists to problematic aspects of treatment as it evolves (Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert, 2007; Wampold, 2015). Implementing ROM in mental health care could have a considerable impact on improving treatment results and preventing failures. Several meta-analyses (Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Lambert & Shimokawa, 2011; Shimokawa, Lambert, & Smart, 2010) support its use. However, the existing body of evidence was found insufficient in a recent Cochrane review (Kendrick et al., 2016) and research results have varied across treatment settings, therapists, and clients. The present randomized controlled trial (RCT) investigated the effects of a ROM system, the *Partners for Change Outcome Management System* (PCOMS; Duncan & Reese, 2015; Miller, Duncan, Sorrell, & Brown, 2005), on therapy outcomes over a four-year implementation period in a hospital-based mental health clinic.

To date, there are ten published randomized trials (RCTs) of PCOMS. Their results are mixed. Reese, Norsworthy, and Rowlands (2009) reported the highest effect sizes (ESs) for PCOMS (d = 0.54 and 0.49) from individual therapy trials at a university counseling center and a graduate training clinic, respectively. Two couple therapy trials (Anker, Duncan, & Sparks, 2009; Reese, Toland, Slone, & Norsworthy, 2010) and a group therapy study at a university counseling center (Slone, Reese, Mathews-Duvall, & Kodet, 2015) also reported moderate ESs in favor of PCOMS. In a study of group substance abuse treatment for soldiers (Schuman, Slone, Reese, & Duncan, 2015) the ES in favor of PCOMS was lower (d = 0.28). No significant effects were found in three RCTs, two of which (Rise, Eriksen, Grimstad, & Steinsbekk, 2016; van Oenen et al., 2016) investigated PCOMS in psychiatric individual treatment settings and the third (Davidsen et al., 2017), in group therapy for eating disorders.

[1] Other terms for ROM are client feedback systems, Patient-Reported Outcome Measures (PROMS), and Feedback-Informed Treatment (FIT).

Finally, in an individual treatment trial at a university counseling center (Murphy, Rashleigh, & Timulak, 2012), the effects of PCOMS depended on what problems the clients presented with: Those with anxiety benefitted from the intervention while those suffering from depression or problems with relationships did not. Similarly, a longitudinal trial with a non-equivalent control group design (Janse, De Jong, Van Dijk, Hutschemaekers, & Verbraak, 2017) found that PCOMS improved individual therapy outcomes for clients with mood disorders, but not for clients with anxiety, somatoform or adjustment disorders.

Understanding when ROM improves outcomes and when it does not is crucial for maximizing its benefits. Some of the variability between studies is likely due to the quality of implementation or the actual use of ROM. There is strong empirical support that the level of implementation influences outcomes of behavior interventions (Durlak & DuPre, 2008), and within ROM research and practice, clinical implementation is acknowledged as challenging, but of vital importance (Boswell, Kraus, Miller, & Lambert, 2015; Lutz et al., 2015; Miller, Hubble, Chow, & Seidel, 2015; Wampold, 2015; Wolpert, 2014).

One aspect of implementation is the fit between ROM and the structure in which the treatment occurs. For instance, Davidsen et al. (2017) observed that therapists in their study had limited flexibility to adjust treatment according to feedback due to the standardized group-format treatment they delivered, which may explain why PCOMS did not improve outcomes in this trial. Likewise, Krägeloh, Czuba, Billington, Kersten, and Siegert (2015) reviewed 27 ROM studies and found that effects were higher when therapists had opportunities to discuss ROM feedback with their clients and allow it to inform the treatment. As noted by de Jong (2016) however, not all therapists use the feedback they receive (e.g. de Jong, van Sluis, Nugter, Heiser, & Spinhoven, 2012; Simon, Lambert, Harris, Busath, & Vazquez, 2012), which suggests that implementation efforts should also target therapists' attitudes, motivation and skills, e.g. through training and supervision. Notably, in only three

PCOMS studies (Anker et al., 2009; Janse et al., 2017; van Oenen et al., 2016) were therapists trained or supervised with some regularity. In the remaining studies, therapists were offered only pre-trial training sessions with no follow-up throughout the trials. Clinical experience suggests that this may be insufficient to achieve the full effects of the intervention; the effective and sustainable use of ROM may require systematic efforts over extended periods of time, often several years (Boswell et al., 2015; Fixsen, Blase, Naoom, & Wallace, 2009; Mellor-Clark, Cross, Macdonald, & Skjulsvik, 2016; Miller et al., 2015).

If the level of implementation influences therapy outcomes and successful implementation takes time and effort to accomplish, we would expect the effects of ROM to increase throughout a period of systematic ROM implementation. Consistent with this, treatment effects were shown to increase in two uncontrolled case studies of clinics where therapists were regularly trained and supervised in the use of ROM (Goldberg, Babins-Wagner, et al., 2016; Miller, Duncan, Brown, Sorrell, & Chalk, 2006). In contrast, therapists' outcomes diminished slightly over time in a longitudinal study where no such continued implementation efforts were performed (Goldberg, Rousmaniere, et al., 2016). To our knowledge, only one controlled trial (Davidsen et al., 2017) has investigated whether the effects of the PCOMS were the same throughout the study period. In this trial, no systematic implementation efforts over time were described, and the results did not differ in the first and second phases of the trial.

Another possibility is that ROM influences outcomes differentially according to some characteristics of clients and/or therapists. As detailed above, two PCOMS trials found differential effects according clients' presenting problems. Furthermore, only the three studies that reported null findings were conducted in psychiatric settings. Consistent with this, Davidson, Perry, and Bell (2015) reviewed 10 studies and observed that ROM effects were lesser in those conducted in severely impaired populations. Regarding therapist effects,

several studies (e.g. Anker et al., 2009; de Jong et al., 2012; Lutz et al., 2015; Simon et al., 2012) have documented that some therapists benefit more than others from working with ROM.

The variability in ROM effects between studies may also be related to how treatment outcomes were assessed. In all of the studies that reported superior treatment effects with PCOMS, outcomes were assessed with PCOMS' progress measure, the Outcome Rating Scale (ORS; Miller, Duncan, Brown, Sparks, & Claud, 2003). In contrast, three studies (Rise et al., 2016; van Oenen et al., 2016; Davidsen et al., 2017) utilized independent outcome measures and found no differences between conditions. Using a ROM progress measure to assess the effects of that same intervention is not uncommon, but somewhat problematic, for several reasons. A bias is introduced if clients in the experimental condition complete the measure in every session and consequently, become more familiar with it than clients in the control condition. The external and internal validity of the findings may be questioned; what exactly is the change that is being measured, and how reliably is it measured? The ORS is an ultra-brief, general, four-item wellbeing scale developed for use as a clinical tool rather than as a research instrument, and more comprehensive measures typically have better psychometric properties (Miller et al., 2003). Clearly, more ROM studies that utilize independent measures are warranted.

The present RCT examined the effects of the ROM intervention PCOMS on therapy outcomes as assessed with an independent measure of symptoms and functioning. The trial took place at a hospital-based mental health clinic in the course of a four-year implementation period, during which time the therapists were regularly trained and supervised in the use of PCOMS. The main hypothesis was that clients receiving treatment with ROM would have better treatment outcomes than clients receiving treatment without ROM. We also hypothesized that implementation time (i.e., the timing of the treatment

within the implementation period) would be positively associated with treatment outcome for ROM cases, that initial severity would be negatively associated with the effect of ROM, and that therapists would differ in terms of the effect of ROM on their clients.

Methods

Design and Randomization

In a naturalistic randomized clinical superiority trial, waitlist psychotherapy clients were randomly assigned to one of two conditions: 'Routine Outcomes Monitoring (ROM)' or 'Treatment As Usual (TAU)'. Randomization was performed by the first or second author using a web-based randomization program for medical research (www.webcrf.medisin.ntnu.no) and a 1:1 allocation ratio. The randomization took place at a different location than that in which the participants received treatment. It was not practically feasible to blind researchers, participants, or therapists to the results of randomization.

Study Setting

The trial was conducted in a general psychiatric outpatient department at a Norwegian hospital-based mental health clinic. The clinic is part of the public Norwegian specialist mental health care system and serves a population of adult clients (age 18 years or older) who suffer from mental health problems of all diagnostic categories. Clients are referred by general practitioners or by other specialist mental health care facilities.

Participants

Clients. The sample consisted of 161 clients who were accepted for treatment at this clinic. Clients were excluded from participation only if unable to complete questionnaires due to illiteracy, very low cognitive functioning, or poor understanding of the Norwegian language. Table 1 shows the demographic information and psychiatric diagnoses of the

participants. The majority of the participants were female, and the mean age was 34 years. Half of the sample did not work due to unemployment, retirement, or being on sick leave. Close to half of the participants were single. The most frequent therapist-assessed diagnoses according to the International Statistical Classification of Diseases and Health Related Problems (ICD-10; World Health Organization, 1992) were affective and anxiety disorders, followed by hyperkinetic disorders (ADHD) and other disorders.

Therapists. All of the therapists on the treatment team were required to treat the participants in this study. A total of 20 therapists (16 women and four men) participated in the study, each treating 1–19 clients (*mean* = 7.6, *SD* = 5.6). Of these, 11 were clinical psychologists, six were psychiatrists, and three were other mental health care professionals. Three therapists had less than 5 years of mental health work experience, two therapists had 5–9 years of experience, seven therapists had 10–15 years of experience, and eight therapists had more than 15 years of experience. On a 7-point Likert scale (1 = *very little* to 7 = *very much*), the therapists reported being most influenced by psychodynamic therapy models (*median* = 6, *range* = 2–7), followed by humanistic/existential (*median* = 5, *range* = 1–6) and cognitive (*median* = 4, *range* = 2–7) models. Due to staff turnover in the study period, their experience working with PCOMS ranged from 1 month to 5 years at the end of the inclusion period. Seven therapists worked at the clinic throughout the trial period. These treated 93 clients (57.8%) from the total sample, of which 66 (62.3%) were included in the analyses (see participant flow and Figure 1 below).

Conditions

Treatment as usual (TAU) condition. Participants in the control condition were given non-manualized outpatient individual psychotherapy. Following an initial assessment phase, therapists and clients determined the treatment focus, approach, and interventions

together. All cases were presented and discussed in interdisciplinary teams during the treatment process. Clients met with their therapist weekly or bi-weekly. The mean number of treatment sessions in the TAU condition was 13.01 (*SD* = 10.92, median = 10, range = 1–54).

Routine Outcomes Monitoring (ROM) condition. Participants in the experimental condition were given the same standard outpatient individual psychotherapy as TAU clients; the only difference was the addition of the PCOMS measures for ROM participants. ROM clients attended a mean of 12.04 sessions (*SD* = 9.35, median = 9, range = 1–45).

In accordance with The International Center for Clinical Excellence (ICCE) Manuals on Feedback-Informed Treatment (Bertolino & Miller, 2012), therapists administered the Outcome Rating Scale (ORS; Miller et al., 2003) and the Session Rating Scale (SRS; Duncan et al., 2003) during the first and last few minutes of every therapy session. The ORS is a four-item measure of wellbeing in different areas, namely symptoms, relational functioning, social role functioning, and global functioning. Similarly, the SRS measures the therapeutic alliance in four items: Therapeutic relationship, goals and topics, approach or method, and overall experience of the alliance. Both measures are scored on visual analogue scales; clients place marks on 10-cm lines that range from poor to good. Numerical values are found by measuring the position of each mark in cm and adding them up, resulting in item scores that range from 0 (minimum wellbeing/experienced the alliance as very poor at today's session) to 10 (maximum wellbeing/experienced the alliance as very good at today's session) and total scores ranging from 0 to 40.

A web-based scoring program (www.fit-outcomes.com) was used to administer the intervention. Participants scored the questionnaires on computer tablets, and their treatment responses from session to session were displayed as graphs and compared to their scores from previous sessions as well as their expected trajectories of change. These were calculated from first-session ORS scores using algorithms provided by Miller and Duncan (2004), and

displayed in the graph together with the clients' actual scores, making deviations immediately visible. When the ORS scores fell below the expected treatment trajectory or when SRS scores fell below the clinical cutoff or dropped by 1 point, warnings were given in the form of yellow or red signs on the screen, depending on how much the actual score diverged from the expected score.

Therapists were trained to share and discuss information gained through the ORS and the SRS with the client. If problems in a client's response to therapy were indicated, therapists were instructed to engage the client in a dialogue about how therapy could be improved, and to adjust the treatment accordingly.

Clinical implementation, training, and supervision. The process of implementing ROM began about six months prior to the onset of the trial, with one of the developers of the PCOMS, Scott D. Miller, giving a one-day training workshop at the clinic. Each therapist was given a copy of the PCOMS manuals (Bertolino & Miller, 2012). One-day training and group supervision workshops were organized twice each year, and training and supervision sessions were conducted once each month throughout study period. During the training events, the therapists were taught how to introduce, administer, interpret, and integrate PCOMS into therapy. In supervision, client cases were discussed. Participation was obligatory for all therapists, but no attendance records were kept. The principal investigators were responsible for much of the training and supervision, and other experienced supervisors and trainers contributed at intervals throughout the implementation process.

Fidelity. To assess whether the PCOMS measures were administered or withheld according to protocol for the two conditions, therapists rated, at each client's treatment termination, how frequently the ORS and the SRS had been administered in that therapy (rated as: 1 = *every session;* 2 =*some sessions*; 3 = *never*). Data was available for 118 cases. In the TAU condition, the PCOMS measures were reported as *never* administered to 59

clients and *every session* to one client. In the ROM condition, the measured were reportedly administered *every session* to 51 clients, *some sessions* to three clients, and *never* to five clients. Three of the latter cases had missing posttreatment data and were not included in the analyses. These data indicate that the therapists administered the PCOMS measures as intended for all cases but six, three of which were included in the analyses.

Measures

Impairment. The primary outcome was posttreatment level of symptoms and psychosocial functioning, measured at baseline and posttreatment with the Behavior And Symptoms Identification Scale (BASIS-32; Eisen, Wilcox, Leff, Schaefer, & Culhane, 1999). BASIS-32 is a 32-item self-report measure of a broad range of symptoms and problems. Items are rated on a 5-point Likert scale (0 = *no difficulty;* 4 = *extreme difficulty*), generating five subscale scores (relation to self/others, daily living/role functioning, depression/anxiety, impulsive/addictive behavior, and psychosis) and an overall mean score, which was utilized for this study. The internal consistency was high, with a Cronbach's alpha of .94 for the pretreatment scores, similar to an earlier report (Eisen et al., 1999). The BASIS-32 was found previously to be sensitive to change and moderately correlated with other measures of symptoms and function (Eisen et al., 1999). Several validation studies (Doerfler, Addis, & Moran, 2002; Hoffmann, Capelli, & Mastrianni, 1997; Jerrell, 2005; Klinkenberg, Cho, & Vieweg, 1998; Russo et al., 1997) have replicated the sound psychometric properties of the BASIS-32.

To help compare our results to those of previous PCOMS studies, we also assessed therapy outcomes with the Outcome Rating Scale (ORS). A pen-and-paper version of the measure was administered to clients in both conditions at baseline and posttreatment (i.e., data were not extracted from the web-based PCOMS database that was used in the

intervention group). The Conditions section (above) describes the item content and scoring of the ORS. Validation studies (Bringhurst & Miller, 2006; Campbell & Hemsley, 2009; Miller & Duncan, 2004) have reported high internal consistency and sensitivity to change in clinical samples, and moderately high concurrent validity with longer measures. In the present trial, the Cronbach's alpha at baseline was .86, and the baseline correlation between the ORS and BASIS-32 was .67. Using the method described by Jacobson and Truax (1991) for determining the clinical significance of change, Miller et al. (2003) established the clinical cutoff separating clinical and non-clinical populations at 25 points. The reliable change index (RCI; the magnitude of change likely to exceed measurement error) was determined to be 5 points (Miller & Duncan, 2004).

The timing of treatment within the implementation period. To investigate whether the effects of ROM on outcomes changed or remained stable over the duration of the trial, we registered for each case the number of months from the beginning of the trial to when the case was initiated, resulting in scores that ranged from 0 (started treatment in November 2012) to 38 (started treatment in January 2016). Figure 1 shows the number of clients who initiated treatment each month of the trial.

Recruitment and Procedure

Inclusion to this trial was performed on a weekly basis. The clinic's intake team, which consisted of health personnel that were not part of the research team, assessed individuals referred for treatment for suitability for treatment and eligibility to participate in this trial. The assessments were based on referral letters, which typically contained a brief description of the presenting problem and relevant medical or psychiatric history. Individuals were considered non-eligible for treatment at the clinic if treatment elsewhere, such as lower level treatment facilities, inpatient or group treatment, was judged to be more suitable. Clients who

were accepted for treatment but excluded from the trial were offered standard outpatient treatment.

Clients were assigned to therapists prior to inclusion in the trial and based on each therapist's current work load, level of experience, and the nature of the client's problems, with more challenging cases being assigned to more experienced therapists. The therapists treated clients in both conditions and the cases were allocated to therapists prior to randomization, minimizing differences between conditions related to therapist experience, competence, treatment models, and case mix.

Individuals deemed eligible for participation were invited to participate via mail and telephone. Prior to their first treatment session, prospective participants met in person with one of the principal investigators to give informed consent, complete baseline measures, and undergo randomization. Shortly thereafter, participants entered treatment.

As is standard routine at the clinic, clients were diagnosed by their therapists during the first few treatment sessions using the M.I.N.I. International Neuropsychiatric Interview (v. 5.0). The diagnoses served clinical purposes only; this trial did not assess the reliability of the diagnostic procedure. All outcome measures were pen-and-paper self-report questionnaires. Baseline measures were completed prior to randomization, and posttreatment measures were mailed to the participants upon treatment termination.

Participant flow. As depicted in Figure 2, the clinic received 1 655 referrals in the trial period. In addition to those who were not considered eligible for treatment at the clinic and thus, participation in the trial, an unknown number of individuals were not invited to participate due to clerical errors (for example, in periods of the trial the intake team forgot to assess all referrals for eligibility). A total of 659 clients (40.0% of all referrals) received invitations to participate. Recruiting ended when 170 individuals (25.9% of those invited to participate) had agreed to participate and been randomized. Nine participants (5.3%) received

no therapy sessions and were discharged without treatment, leaving 161 participants in the final sample. Of these, one participant (0.6% of final sample) had missing baseline measures and 47 (29.2%) had missing posttreatment measures; thus, 113 cases (70.2%) were included in the descriptive outcomes analysis. In addition, seven participants (4.3%) changed therapists during treatment due to staff turnover and consequently, had missing data at level 2, leaving 106 cases (66.5%) for the multilevel analyses.

The first participant started treatment in November 2012 and the last one in January 2016. Data collection for this study was completed in February 2017, resulting in a trial period of about four years.

The study protocol was approved by the Regional Committee for Research Ethics (Case number 2011/1711), and the trial was registered on Clinical Trials (clinicaltrials.gov; identifier: NCT01796223).

Data Analysis

Descriptive therapy outcomes. Cohen's *d* effect sizes for pre-post change in the ROM v. TAU condition were calculated[2] based on scores on the primary outcome measure BASIS-32. To our knowledge, there is no established clinical cutoff or reliable change index (Jacobson & Truax, 1991) for assessing the clinical significance of change on this measure. The clinical cutoff could not be determined due to a lack of norm data, but like Jerrell (2005) and Eisen, Ranganathan, Seal, and Spiro III (2007), we calculated the Reliable Change Index - Improved Difference, RCI_{ID} (Hageman & Arrindell, 1993) as an indication of whether the observed differences between pre- and posttreatment scores on the BASIS-32 were likely to exceed measurement error and thus represent real change. The observed difference score was adjusted for regression to the mean and then expressed in terms of the standard error of the

[2] Formula for pre-post change effect size: $d = (MEAN_{\text{pre-post difference ROM}} - MEAN_{\text{pre-post difference TAU}}) / SD_{\text{pooled,}}$ where $SD_{\text{pooled}} = (SD^2_{\text{pre}} + SD^2_{\text{post}})/2$

sum of 2 independent (measurement) error components (i.e., the square root of the sum of pre- and post-measurement error variance). Changes exceeding 1.96 standard errors in either direction, which corresponded to ± 0.55 raw score points on the BASIS-32 in this data set, were classified as indicating reliable change. Cohen's *d* pre-post effect size as well as classification according to the RCI and clinically significant change is also presented for the ORS. We did not apply null hypothesis test statistics at this stage of the analysis as the assumption of independence of observations was violated by the potential shared covariance by clients treated by the same therapist.

Tests of hypotheses. To properly model the nested structure with each therapist treating several clients (Wampold & Serlin, 2000), a series of multilevel models (MLMs; Snijders & Bosker, 2012) were fitted using the Mplus statistical software (Muthén & Muthén, 1998-2017), with maximum likelihood estimation. Model fit was assessed by Akaike (AIC) and Bayesian (BIC) estimations and the loglikelihood (llg) chi square test. Total variance explained by each model was assessed by R^2, and the proportion of variance explained by differences between therapists, by the intraclass correlation coefficient (ICC)[3].

First, within-therapist and between-therapist variability in posttreatment BASIS-32 scores were parsed in an unconditional or null model (i.e. no predictors). Models 1 through 4 then examined the effects of client level variables in random intercept, fixed slope models, allowing for therapist variability in the intercepts for each predictor but modeling slope coefficients as being equal across therapists. We first controlled for clients' grand mean centered pretreatment BASIS-32 score (model 1) and examined the overall effects of ROM (model 2). We then tested for moderation of the ROM effect by the point in time within the

[3] Formula for the intraclass correlation coefficient: ICC = $\tau_{00}/(\tau_{00} + \sigma^2)$, where τ_{00} is the between-therapist variability in the dependent variable and σ^2, the within-therapist variability (Raudenbusch & Bryk, 2002; see also Adelson & Owen, 2012).

implementation period at which clients were treated (model 3) and clients' initial impairment (model 4). Model 4 was expressed at the client level by the equation

$$Y_{ij} = b_{0\,j} + b_{1j}(\text{Pre}_{ij}) + b_{2j}(\text{Condition}_{ij}) + b_{3j}\,(\text{Time}_{ij}) + b_{4j}(\text{Time}_{ij})(\text{Condition}_{ij}) + b_{5j}(\text{Pre}_{ij})(\text{Condition}_{ij}) + e_{ij}$$

where Y_{ij} is the posttest score for client i treated by therapist j; b_{0j} is the intercept for therapist j, b_{1j} is the estimate for the pretreatment distress slope; b_{2j} for the condition slope; b_{3j} for the time slope; b_{4j} for the condition by time interaction slope; b_{5j} for the condition by pretreatment distress slope, and e_{ij} is the residual for client i treated by therapist j. At the therapist level, the intercept b_0 for therapist j is expressed by the equation

$$b_{0j} = \gamma_{00} + \mu_{0j}$$

where γ_{00} is the mean intercept and μ_{0j} is the residual for the intercept for therapist j. Finally, models 5a, b, and c investigated therapist variability in ROM effects by retaining all predictors from model 4 and fitting random slopes between levels for Condition (model 5a), the Time*Condition interaction (model 5b), and the Pre*Condition interaction (model 5c). We did not include any therapist level predictors.

Like in many other naturalistic studies, missing data were a challenge in this RCT. To assess whether data were missing completely at random (MCAR) or depended upon the observed variables (MAR), we compared cases with complete data sets to cases with missing posttreatment measures on all baseline variables. With only two measurement points we were not able to use more sophisticated methods like pattern-mixture models to detect patterns of missingness and to assess whether data were missing not at random (MNAR). Missing values were imputed using the maximum likelihood imputation procedure (Schafer & Graham, 2002) with all other observed variables as auxiliary variables. All analyses were performed twice, first with complete cases only (i.e. no imputation) and then with the imputed data set.

In the Results section, the analyses of cases with complete data sets are presented in all tables and analyses of the imputed data set are described in the text.

The a priori hypotheses in this trial were directional (i.e. superior outcomes in the ROM condition, positive effects of initial impairment on posttreatment distress levels, and negative effects of clients' time of treatment). Accordingly, we report one-sided significance tests with an alpha level of .05. For the corresponding one-sided confidence intervals (CIs) we report 90% CIs; here, the upper bound represent the value below which we would expect 95 % of future observations to fall (i.e. the error rate is 5% because only one side of the CIs are of interest when hypotheses are directional. In contrast, when testing the null hypothesis that conditions are equal, both sides of the interval are taken into account and consequently, a 90% CI gives an error rate of 10%; Pocock, 2003).

Results

Descriptive Analyses

Potential sources of bias. There were no significant differences in baseline distress between conditions as measured by the BASIS-32 and the ORS (Table 1). Diagnoses were similarly distributed across conditions, as were the proportions of undiagnosed clients and clients with comorbid disorders (*p*s >.05 for all). No statistically significant differences were found between ROM and TAU clients for demographic variables except for social network; significantly more clients in the ROM than TAU condition reported having nobody in whom they could confide ($\chi^2(1) = 4.826$, $p = .028$). There was no significant difference in number of sessions attended in the TAU v. ROM conditions. The conditions were similar in number of cases with missing posttreatment measures (TAU: $n = 24$, 28.9%; ROM: $n = 23$, 29.5%; $\chi^2(1) = .008$, $p = .30$). Comparing cases with missing posttreatment measures to those with complete data sets, we found no differences in the baseline BASIS-32 and ORS scores or in

the distribution of diagnoses ($p > .05$ for all), but clients with missing posttreatment measures were more often single (70.2% v. 36.6%; $\chi^2(1) = 15.027$, $p < .001$) and living alone (27.7% v. 9.8%; $\chi^2(1) = 8.220$, $p = .004$), indicating that the data were not missing completely at random (MCAR).

To examine if the variables that differed between the conditions (i.e. social network) and between the cases missing data and those with complete data sets (i.e. marital status and living situation) biased our models, we examined whether these variables predicted posttreatment BASIS-32 scores while controlling for pretreatment scores. This was only the case for marital status ($b = -0.315$, $SE = 0.117$, $p = .007$). Adding these variables to the models described below did not result in major changes to the estimates for the other predictors or in significantly better model fit, and explained little additional variance (0.2%–1.5%). Thus, we did not control for any of these variables in the final models.

To examine whether therapist turnover influenced the estimate for the effect of treatment timing on therapy outcomes, we fitted model 3 (see data analysis section) to a subsample of cases ($n = 66$) that were treated by the therapists who were employed throughout the study. The estimate for the timing by condition interaction was comparable to the full sample ($b = -0.013$) but non-significant, possibly due to the loss of statistical power from excluding approximately one third of the cases.

Descriptive therapy outcomes. Table 2 shows the mean BASIS-32 pre- and posttreatment scores for all clients with complete data sets (i.e. no imputation and including the six cases that changed therapist mid-treatment). Clients in the ROM condition demonstrated greater pre- to posttreatment improvement than those in the TAU condition. The pre-post ES Cohen's *d* for treatment in the ROM v. TAU condition was 0.42. More clients were classified as improved according to the RCI in the ROM than TAU condition

(Table 3), with an odds ratio (OR) for achieving reliable improvement for ROM clients of 2.45.

Mean pretreatment scores on the ORS were 14.62 (*SD* = 7.69) and 14.78 (*SD* = 9.04) in the ROM and TAU condition, respectively, and mean posttreatment scores were 26.42 (*SD* = 10.54) and 22.26 (*SD* = 10.63). The pre-post Cohen's *d* for treatment in the ROM v. TAU condition was 0.42. More clients in the ROM than TAU condition were classified as having reliable improvement (ROM: *n* = 40, 74.1%; TAU: *n* = 32, 58.2%; OR = 2.05) and clinically significant improvement (ROM: *n* =26, 48.1%; TAU: *n* = 18, 32.7%; OR = 1.91).

Tests of Hypotheses

Effect of ROM on therapy outcomes. In model 2 (Table 4), the main effect of ROM on outcomes was examined by adding Condition as a predictor to the base model. The negative and significant coefficient for Condition (*b* = -0.180, *SE* = 0.101, 90% CI [-0.346, -0.014], *p* = .037) indicates less posttreatment distress for clients in the ROM condition than those in the TAU condition. The standardized effect size *d* for ROM, calculated by dividing the coefficient for ROM by the pooled standard deviation, was 0.261. Model 2 explained 25.8% of the variance in outcomes, 9.7% of which was accounted for by therapists, as indicated by the ICC. The AIC for this model indicated better model fit than the previous base model while the BIC indicated poorer model fit; the loglikelihood chi square test was non-significant ($\chi^2(1) = 2.712$, $p = .100$) indicating that the difference in model fit was not statistically significant. After imputation of missing posttreatment values for the 47 cases with missing posttreatment data, the coefficient for Condition was slightly larger (*b* = -0.206, *SE* = 0.077, 90% CI [-0.332, -0.079], *p* = .004).

Moderation of the ROM effect by implementation time. In model 3, the predictors Time and the Time*Condition interaction was added to the previous model. As hypothesized,

the Time*Condition interaction, which modeled the extent to which the timing effect was greater in the ROM (coded 1) than TAU (coded 0) condition, negatively and significantly predicted outcomes (b = -0.018, SE = 0.010, 90% CI [-0.034, -0.001], p = .036); for each month of the trial period ROM clients' posttreatment distress levels diminished by 0.018 points more on the BASIS-32 than the TAU clients, i.e. the difference in outcomes between conditions increased over time. Specifically, the contrasted per-month rates for ROM and TAU were b = -0.008 for ROM and b = 0.010 for TAU. Thus, ROM use produced better outcomes over time, whereas the outcomes in TAU decreased slightly over time. The estimate for change over time in the TAU condition (i.e. the coefficient for Time in Table 4) was non-significant and the corresponding estimate in the ROM condition was significant (p = .029; obtained by reversing the coding so that ROM clients were coded 0 for Condition and the estimate for Time represented the time effect for ROM cases only). Model 3 explained 28.2% of the variance in outcomes, and therapists accounted for 9.4% of the variance. The model fit indices AIC and BIC were both higher than the previous model indicating poorer model fit, but this difference in model fit was not statistically significant as indicated by the loglikelihood chi square test ($\chi^2(2) = 3.222$, p =.200). Repeating the analysis with the imputed data set resulted in a Time*condition coefficient of -0.015 (SE = 0.007, 90% CI [-0.029, -0.004], p = .014).

To illustrate the magnitude of the change in ROM effects over time, we estimated posttreatment BASIS-32 scores for clients with average initial distress (Pre = 0) in the ROM condition (Condition = 1) using the regression equation from model 3 (i.e. $\text{BASIS-32}_{\text{POST}}$ = 0.829 + 0.537*Pre + 0.171*Condition + 0.010*Time - 0.018*Time*Condition). Clients initiating treatment in the beginning of the study (Time = 0) had an estimated posttreatment score of 1.000 and clients initiating treatment in the last month (Time = 38), 0.696. This corresponds to d = 0.743 in month 0 and d = 1.210 in month 38 (based on $MEAN_{\text{PRE ROM}}$

=1.484 and $SD_{\text{POOLED ROM}}$ = 0.651 for the cases included in the multilevel models). That is, there was an estimated effect size growth of 0.467 from the first to the last month of the study for clients with average initial distress in the ROM condition.

Moderation of the ROM effect by initial impairment. In model 4, a parameter for the Pre*Condition interaction was added to the previous model. As shown in Table 4, the Pre*Condition interaction had a negative, non-significant effect on outcome (*b* = -0.092, *SE* =0 .151, 90% CI [-0.389, 0.158], *p* = .275), indicating that the ROM effect did not differ according to initial impairment. The model fit was poorer than the previous model as indicated by higher AIC and BIC scores but the loglikelihood chi square test indicated that this difference was not statistically significant ($\chi^2(1) = 0.274$, *p* = .601). The model explained 28.2% of the variance in outcomes, 9.7% of which was due to therapists. After imputation of missing values, the Pre*Condition coefficient was -0.004 (*SE* = 0.083, 90% CI [-0.141, 0.133], *p* = .482).

Therapist variability. The ICCs in models 1 through 4 indicate that therapists accounted for 9%–10% of the variance in outcomes. Fitting random slopes between levels to investigate whether ROM influenced outcomes differently for different therapists resulted in the following random slope coefficients: For Condition (model 5a), 0.161 (*SE* = 0.232, 90% CI [-0.221, 0.542], *p* = .489); for the Time*Condition interaction (model 5b), -0.015 (*SE* =0 .011, 90% CI [-0.033, 0.002], *p* = .146), and for the Pre*Condition interaction (model 5c), -0.204 (*SE* = 0.171, 90% CI [-0.486, 0.077], *p* = .232). Repeating the analyses on the imputed dataset similarly produced non-significant random slopes for all predictors and interactions. Thus, there were no significant differences between therapists regarding ROM effects.

Discussion

The results of this naturalistic randomized clinical trial supported the main hypothesis that clients receiving treatment with the ROM system PCOMS had superior outcomes to those receiving treatment as usual. Clients in the ROM condition were 2.5 times more likely to demonstrate reliable improvement in their symptoms and functioning as measured by the BASIS-32 (Eisen et al., 1999). The advantage of ROM over TAU remained when controlling for therapist variability and clients' pretreatment distress levels, with an estimated overall effect size of ROM vs. TAU of $d = 0.26$, which is considered a small effect size (Cohen, 1988). However, the effects of ROM differed according to the timing of clients' treatment within the implementation period. For each month in the four-year long trial, the difference in posttreatment distress between conditions was estimated to increase by 0.018 points on the BASIS-32, which corresponded to a growth in effect size of $d = 0.47$ from the first to the last month of the trial for clients in the ROM condition with average pretreatment distress levels (specifically, the estimated Cohen's *d* for treatment in the first month of the trial was 0.74 and for treatment in the last month, 1.21). The effects of ROM did not depend on how impaired clients were at intake. Differences between therapists accounted for 9%–10% of the variability in outcomes across both conditions, but there were no significant differences between therapists in how ROM influenced outcomes.

The increasing superiority of ROM compared to TAU over time was due to a significant improvement in the ROM condition, and a corresponding non-significant deterioration in the TAU condition. To our knowledge, this is the first RCT to show increases in ROM effects over time, although two uncontrolled cases studies (Goldberg, Babins-Wagner et al., 2016; Miller et al., 2006) have demonstrated increasing treatment effects in clinics where therapists work with ROM. We did not measure any implementation factors and consequently, the interpretation of this finding is uncertain. One explanation is that

PCOMS was used increasingly more effectively over time, possibly as a result of the continued implementation efforts that took place over the duration of the trial. In support of this, it is well established that higher levels of implementation are associated with better outcomes for behavior interventions (Durlak & DuPre, 2008).

The main implementation strategy utilized in this study was to provide regular and frequent ROM training and supervision throughout the trial period. We organized biannual one-day workshops as well as monthly ROM meetings, which amounts to more training than that described in other controlled studies of the PCOMS. Comparing our study to that of Goldberg, Babins-Wagner, et al. (2016), we note some similarities in how ROM was taught and supervised. In both clinics, the principles delineated in the ICCE Manuals in Feedback-Informed Treatment (Bertolino & Miller, 2012) as well as the recommendations of Miller, Hubble, Chow, and Seidel (2013) were followed. We prioritized discussing specific non-progressing cases, emphasizing the client's experience of the working alliance rather than diagnoses and therapeutic models or techniques. Therapists were encouraged to deliberately practice their therapeutic skills based on their clients' feedback. An overall goal was to foster a 'culture for feedback' (Bertolino & Miller, 2012) that valued clients' feedback and critical evaluation of our own practices. Interestingly, Goldberg, Rousmaniere et al. (2016) found treatment effects to diminish slightly over seven years in a clinic where therapists used ROM but no training or supervision on a regular basis was reported. Together these studies indicate that simply using ROM may be insufficient improve treatment effects over time, and that an ongoing shared focus on how to improve outcomes with the use of ROM may be crucial.

Beyond training and supervision, some basic practical and structural factors may have facilitated the implementation of PCOMS in this clinic. Unlike the group therapy setting investigated by Davidsen et al. (2017), the treatment context was flexible enough to allow therapists to adjust treatment according to the ROM feedback. The project was initiated and

run by therapists who used ROM as a tool in their clinical work rather than as a quality control system. The latter has been suggested as a potential cause for iatrogenic effects of these interventions (Wolpert, 2014). Others (Boswell et al., 2015; Gleacher et al., 2016) have emphasized the importance of a supportive clinic management and financial resources. In our study, management nurtured the initiative with moral and financial support, and additional resources were provided by external funding. Thus, we were able to purchase computer tablets and software for administering the PCOMS measures, which have been noted as important for the successful use of ROM (Bickman et al., 2016; Lucock et al., 2015). We also engaged highly competent trainers and supervisors, and the principal investigators were given time away from their clinical work to provide on-site training, supervision, and support. This implementation strategy has been recommended by several authors (Bickman et al., 2016; Boswell et al., 2015; Fixsen et al., 2009; Mellor-Clark et al., 2016).

The effect size (ES) for ROM in the present study was the lowest of any of the PCOMS trials that found the intervention to improve outcomes. Several explanations are possible for this. It is likely that the low overall ES was related to the growth of PCOMS' effects over time. Since the intervention had little influence on outcomes at the beginning of the trial, it is not surprising that the ROM estimate was modest when the timing of treatment was unaccounted for. In addition, this was the first trial to find superior effects for PCOMS in a psychiatric clinic, and the low ES is consistent with other ROM studies targeting impaired populations (Davidson et al. 2015).

Therapist variability may also explain the low ES for ROM in this study. The proportion of variability in therapy outcomes accounted for by therapists (9%–10%) was slightly larger than that commonly reported in psychotherapy studies (Baldwin & Imel, 2013), indicating that working with ROM did not reduce the variability between therapists in this sample. Most PCOMS studies have neglected to control for the covariance of clients

treated by the same therapists in multilevel models, which may compromise standard errors and inflate effect sizes (Adelson & Owen, 2012); exceptions include Anker et al. (2009) and Reese et al. (2010), both of which reported moderately high ES in favor of PCOMS but also lower ICCs for therapists (.04 and .02, respectively) than in the present study. In the present trial, there were no significant differences between therapists in ROM effects, but like in many other clinical studies, few therapists were included in the sample, which may have limited the statistical power to detect differences between therapists.

Although Anker et al. (2009) did find PCOMS to be effective also as assessed by their secondary outcome measure Locke-Wallace Marital Adjustment Test (Locke & Wallace, 1959), this is the first study to demonstrate effects of PCOMS on a different main outcome measure than the PCOMS' progress measure ORS. Consequently, a hypothesis for the low ES in this trial is that differences between clients in treatment with and without PCOMS are more difficult detect on independent outcome measures. This hypothesis is however not supported by our results; we calculated pre-post Cohen's *d* as well as odds ratios for reliable and clinically significant change from both ORS and BASIS-32 scores and found the results to be very similar, suggesting that the low ES was not related to how outcomes were assessed in this study.

Limitations

As this trial was conducted within the daily practice in a clinical setting, experimental control was challenging. Due to a failure to assess all referrals for eligibility, difficulties recruiting those who were eligible and missing data for those who were included in the trial, data were only available for analysis from a small proportion of all prospective participants. Clients with complete data sets and clients with missing data points were similar on most baseline variables, but it was not possible to assess if they differed in terms of therapy

process or outcome variables. Thus, the extent to which our results are generalizable to other populations is somewhat unclear. Another possible limitation to the external validity of this study is the use of highly trained experts, time spent on implementation, and use of computer scorings, which may not be equally accessible in all treatment settings.

Limitations in the design of this study make it difficult to assess the extent to which our results were influenced by confounding variables. First, the interpretation of the changes in ROM effects over time is uncertain due to the lack of measures of other implementation variables. Second, the specific type or dose of psychotherapy, medical treatment, or treatment outside the clinic was not assessed. Third, both of the principal researchers were advocates for ROM, and the therapists were supervised by one of the developer of PCOMS during parts of the trial. Fourth, although the therapists' self-reported administration of the PCOMS measures indicated high fidelity to the intervention, we did not assess other aspects of adherence to the PCOMS manual, nor did we measure the therapists' attitudes towards ROM. Fifth, we did not assess if, and to what extent, ROM cases were discussed during the supervision sessions. Seventh, about one third of the client sample was treated by therapists who worked at the clinic for parts of the trial period. The examination of therapist variability did not reveal significant differences between therapists in ROM effects, but suffered from low statistical power at level 2. However, including only the cases treated by therapists who worked at the clinic throughout the trial in the model resulted in a similar, although slightly smaller, estimate was obtained for the timing effect. This suggests that turnover did not substantially bias our results.

Implications and Directions for Future Research

Our results support the use of PCOMS in the outpatient treatment of adults. The demonstration of increases in ROM effects over time suggests that ROM may be less

effective at the onset of an implementation process compared to later on, and we find that our results make a case for sustained implementation efforts.

Additional research is needed to further explore the nuances of when ROM is and is not effective. Future studies should measure moderators at the client, therapist, and clinic levels as well as potential mediators. More research is needed on how to best achieve successful implementation. When researching ROM in practice settings, researchers should be aware that the effects might change over time and should thus consider measurement timing and assess how well the intervention is implemented. Finally, it is our hope that future studies will be able to investigate therapist effects in ROM with sufficiently large samples.

References

Adelson, J. L., & Owen, J. (2012). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy, 49*(2), 152-162.

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77*(4), 693-704.

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 256-297). Hoboken, NJ: John Wiley & Sons.

Bertolino, B., & Miller, S. D. (2012). *ICCE manuals on feedback-informed treatment.* (Vol. 1-6). Chicago, IL: ICCE Press.

Bickman, L., Douglas, S. R., De Andrade, A. R. V., Tomlinson, M., Gleacher, A., Olin, S., & Hoagwood, K. (2016). Implementing a measurement feedback system: A tale of two sites. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 410-425. doi:10.1007/s10488-015-0647-8

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19.

Bringhurst, M. D. L., & Miller, S. D. (2006). The Reliability and Validity of the Outcome Rating Scale: A Replication Study of a Brief Clinical Measure. *Journal of Brief Therapy, 5*(1). 23-30.

Campbell, A., & Hemsley, S. (2009). Outcome Rating Scale and Session Rating Scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist, 13*(1), 1-9.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillside, NJ, England: Lawrence Earlbaum Associates.

Davidsen, A., Poulsen, S., Lindschou, J., Winkel, P., Tróndarson, M., Waaddegaard, M., & Lau, M. (2017). Feedback in group psychotherapy for eating disorders: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, 85(5), 484-494. doi:10.1037/ccp0000173.

Davidson, K., Perry, A., & Bell, L. (2015). Would continuous feedback of patient's clinical outcomes to practitioners improve NHS psychological therapy services? Critical analysis and assessment of quality of existing studies. *Psychology and Psychotherapy: Theory, Research and Practice, 88*(1), 21-37.

de Jong, K. (2016). Deriving implementation strategies for outcome monitoring feedback from theory, research and practice. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 292-296. doi:10.1007/s10488-014-0589-6

de Jong, K., van Sluis, P., Nugter, M. A., Heiser, W. J., & Spinhoven, P. (2012). Understanding the differential impact of outcome monitoring: Therapist variables that moderate feedback effects in a randomized clinical trial. *Psychotherapy Research, 22*(4), 464-474.

Doerfler, L. A., Addis, M. E., & Moran, P. W. (2002). Evaluating mental health outcomes in an impatient setting: Convergent and divergent validity of the OQ-45 and BASIS-32. *The Journal of Behavioural Health Services & Research, 29*(4), 394-403.

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a “working” alliance measure. *Journal of Brief Therapy, 3*(1), 3-12.

Duncan, B. L., & Reese, R. J. (2015). The Partners for Change Outcome Management System (PCOMS): Revisiting the client's frame of reference. *Psychotherapy, 52*(4), 391-401.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology, 41*(3-4), 327-340.

Eisen, S. V., Ranganathan, G., Seal, P., & Spiro III, A. (2007). Measuring clinically meaningful change following mental health treatment. *The Journal of Behavioural Health Services & Research, 34*(3), 272-289.

Eisen, S. V., Wilcox, M., Leff, H. S., Schaefer, E., & Culhane, M. A. (1999). Assessing behavioral health outcomes in outpatient programs: Reliability and validity of the BASIS-32. *The Journal of Behavioural Health Services & Research, 26*(1), 5-17.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19*(5), 531-540.

Gleacher, A. A., Olin, S. S., Nadeem, E., Pollock, M., Ringle, V., Bickman, L., . . . Hoagwood, K. (2016). Implementing a measurement feedback system in community mental health clinics: A case study of multilevel barriers and facilitators. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 426-440. doi:10.1007/s10488-015-0642-0

Goldberg, S. B., Babins-Wagner, R., Rousmaniere, T., Berzins, S., Hoyt, W. T., Whipple, J. L., . . . Wampold, B. E. (2016). Creating a climate for therapist improvement: A case study of an agency focused on outcomes and deliberate practice. *Psychotherapy, 53*(3), 367-375.

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A

longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63*(1), 1-11.

Hageman, W. J., & Arrindell, W. A. (1993). A further refinement of the reliable change (RC) index by improving the pre-post difference score: Introducing RC ID. *Behaviour research and therapy, 31*(7), 693-700.

Hoffmann, F. L., Capelli, K., & Mastrianni, X. (1997). Measuring treatment outcome for adults and adolescents: Reliability and validity of BASIS-32. *The Journal of Mental Health Administration, 24*(3), 316-331.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*(10), 1059-1064.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19.

Janse, P. D., De Jong, K., Van Dijk, M. K., Hutschemaekers, G. J., & Verbraak, M. J. (2017). Improving the efficiency of cognitive-behavioural therapy by using formal client feedback. *Psychotherapy Research*, 27(5), 525-538. doi:10.1080/10503307.2016.1152408

Jerrell, J. M. (2005). Behavior and Symptom Identification Scale 32. Sensitivity to change over time. *The Journal of Behavioral Health Services and Research, 32*(3), 341-346.

Kendrick, T., El-Gohary, M., Stuart, B., Gilbody, S., Churchill, R., Aiken, L., . . . Moore, M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Database of Systematic Reviews*(7), 1-108. doi:10.1002/14651858.CD011119.pub2

Klinkenberg, W. D., Cho, D. W., & Vieweg, B. (1998). Reliability and validity of the interview and self-report versions of the BASIS-32. *Psychiatric Services, 49*(9), 1229-1231.

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *The British Journal of Psychiatry, 195*(1), 15-22.

Krägeloh, C. U., Czuba, K. J., Billington, D. R., Kersten, P., & Siegert, R. J. (2015). Using feedback from patient-reported outcome measures in mental health services: a scoping study and typology. *Psychiatric Services, 66*(3), 224-241.

Lambert, M. J. (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.

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy, 48*(1), 72-79.

Locke, H. J., & Wallace, K. M. (1959). Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage and family living, 21*(3), 251-255.

Lucock, M., Halstead, J., Leach, C., Barkham, M., Tucker, S., Randal, C., . . . Waters, E. (2015). A mixed-method investigation of patient monitoring and enhanced feedback in routine practice: Barriers and facilitators. *Psychotherapy Research, 25*(6), 633-646.

Lutz, W., Rubel, J., Schiefele, A.-K., Zimmermann, D., Böhnke, J. R., & Wittmann, W. W. (2015). Feedback and therapist effects in the context of treatment outcome and treatment length. *Psychotherapy Research, 25*(6), 647-660.

Mellor-Clark, J., Cross, S., Macdonald, J., & Skjulsvik, T. (2016). Leading horses to water: Lessons from a decade of helping psychological therapy services use routine outcome

measurement to improve practice. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 279-285. doi:10.1007/s10488-014-0587-8

Miller, S. D., & Duncan, B. (2004). *The Outcome and Session Rating Scales: Administration and scoring manual*. Chicago: Institute for the Study of Therapeutic Change.

Miller, S. D., Duncan, B. L., Brown, J., Sorrell, R., & Chalk, M. B. (2006). Using formal client feedback to improve retention and outcome: Making ongoing, real-time assessment feasible. *Journal of Brief Therapy, 5*(1), 5-22.

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J., & Claud, D. (2003). The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy, 2*(2), 91-100.

Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology, 61*(2), 199-208.

Miller, S. D., Hubble, M. A., Chow, D., & Seidel, J. (2015). Beyond measures and monitoring: Realizing the potential of feedback-informed treatment. *Psychotherapy, 52*(4), 449-457.

Miller, S. D., Hubble, M. A., Chow, D. L., & Seidel, J. A. (2013). The outcome of psychotherapy: Yesterday, today, and tomorrow. *Psychotherapy, 50*(1), 88-97.

Murphy, K. P., Rashleigh, C. M., & Timulak, L. (2012). The relationship between progress feedback and therapeutic outcome in student counselling: A randomised control trial. *Counselling Psychology Quarterly, 25*(1), 1-18.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's guide* (Eighth ed.). Los Angeles, CA: Muthén & Muthén.

Pocock, S. J. (2003). The pros and cons of noninferiority trials. *Fundamental & Clincial Pharmacology, 17*(4), 483-490.

Raudenbusch, S., & Bryk, A. (2002). *Hierarchical linear models* (Second ed.). Newbury Park, CA: Sage.

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy: Theory, Research, Practice, Training, 46*(4), 418-431.

Reese, R. J., Toland, M. D., Slone, N. C., & Norsworthy, L. A. (2010). Effect of client feedback on couple psychotherapy outcomes. *Psychotherapy: Theory, Research, Practice, Training, 47*(4), 616-630.

Rise, M. B., Eriksen, L., Grimstad, H., & Steinsbekk, A. (2016). The long-term effect on mental health symptoms and patient activation of using patient feedback scales in mental health out-patient treatment. A randomised controlled trial. *Patient education and counseling, 99*(1), 164-168.

Russo, J., Roy-Byrne, P., Jaffe, C., Ries, R., Dagadakis, C., Dwyer-O'Connor, E., & Reeder, D. (1997). The relationship of patient-administered outcome assessments to quality of life and physician ratings: Validity of the BASIS-32. *The Journal of Mental Health Administration, 24*(2), 200-214.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147-177.

Schuman, D. L., Slone, N. C., Reese, R. J., & Duncan, B. (2015). Efficacy of client feedback in group psychotherapy with soldiers referred for substance abuse treatment. *Psychotherapy Research, 25*(4), 396-407.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*(3), 298-311.

Simon, W., Lambert, M. J., Harris, M. W., Busath, G., & Vazquez, A. (2012). Providing patient progress information and clinical support tools to therapists: Effects on patients at risk of treatment failure. *Psychotherapy Research, 22*(6), 638-647.

Slone, N. C., Reese, R. J., Mathews-Duvall, S., & Kodet, J. (2015). Evaluating the efficacy of client feedback in group psychotherapy. *Group Dynamics: Theory, Research, and Practice, 19*(2), 122-136. doi:10.1037/gdn0000026

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and applied multilevel modeling* (2nd ed.). London: Sage.

van Oenen, F. J., Schipper, S., Van, R., Schoevers, R., Visch, I., Peen, J., & Dekker, J. (2016). Feedback-informed treatment in emergency psychiatry; a randomised controlled trial. *BMC Psychiatry, 16*(1), 110-122. doi:10.1186/s12888-016-0811-z

Wampold, B. E. (2015). Routine outcome monitoring: Coming of age—With the usual developmental challenges. *Psychotherapy, 52*(4), 458-462. doi:10.1037/pst0000037

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological methods, 5*(4), 425-433.

Wolpert, M. (2014). Uses and abuses of Patient Reported Outcome Measures (PROMs): Potential iatrogenic impact of PROMs implementation and how it can be mitigated. *Administration and Policy in Mental Health and Mental Health Services Research, 41*(2), 141-145.

Table 1

Participant characteristics

Variables	Total	Condition		*p*
		TAU	ROM	
Demographic information				
Male, *n* (%)	58 (36.7)	32 (40.0)	26 (33.3)	.385
Age in years, *M (SD)*	34.1 (11.6)	34.6 (12.0)	33.4 (11.3)	.880
Single, *n* (%)	74 (46.5)	42 (51.2)	32 (41.6)	.222
Living alone, *n* (%)	24 (15.1)	12 (14.6)	12 (15.6)	.867
Has nobody to confide in, *n (%)*	28 (17.8)	9 (11.3)	19 (24.7)	.028
No education beyond primary school, *n* (%)	27 (17.1)	13 (16.0)	14 (18.2)	.722
Not working, *n* (%)	77 (50.0)	38 (47.5)	39 (47.3)	.519
ICD-10 diagnoses				.446
Affective disorders, *n* (%)	59 (30.1)	29 (28.4)	30 (31.9)	
Anxiety disorders, *n* (%)	59 (30.1)	29 (28.4)	30 (31.9)	
Hyperkinetic disorders, *n* (%)	20 (10.2)	14 (13.7)	6 (6.4)	
Personality disorders, *n* (%)	17 (8.7)	7 (6.9)	10 (10.6)	
Other, *n* (%)	19 (9.7)	12 (11.8)	7 (7.4)	
Undiagnosed, *n* (%)	22 (11.2)	11 (10.8)	11 (11.7)	
Two diagnoses, *n* (%)	35 (17.9)	19 (18.6)	16 (17.0)	.769

Notes. TAU = treatment as usual condition; ROM = Routine Outcome Monitoring condition; *p* = p-value for the difference between conditions (chi-square test for categorical and t-tests for continuous variables). Diagnostic assessments were performed by therapists and were based on the Mini International Neuropsychiatric Interview (M.I.N.I.). When participants had two diagnoses, both were registered.

Table 2

The mean pre- and posttreatment scores on the Behavior and Symptoms Identification Scale (BASIS-32) for each condition

Time of measurement	TAU *M (SD)*	ROM *M (SD)*
Pretreatment score	1.40 (0.59)	1.53 (0.66)
Posttreatment score	.99 (0.69)	.84 (0.66)

Notes. TAU = treatment as usual condition; ROM = Routine Outcome Management condition.

Table 3

Classification of outcomes based on pre- to posttreatment changes in the Behavior and Symptoms Identification Scale (BASIS-32) scores for each condition

Outcome classification	TAU *n (%)*	ROM *n (%)*
Improved	21 (36.2)	32 (58.2)
No change	32 (55.2)	19 (34.5)
Deterioration	5 (8.6)	4 (7.3)

Notes. TAU = treatment as usual condition; ROM = Routine Outcome Management condition. Cases were classified according to the Reliable Change Index – Improved Difference (RCI_{ID}; Hageman & Arrindell, 1993).

Table 4

Effects of Routine Outcome Monitoring (ROM) and moderating variables on the posttreatment Behavior and Symptoms Identification Scale (BASIS-32) scores

Effects	Model 0	Model 1	Model 2	Model 3	Model 4
Fixed effects					
Intercept	0.929***	0.942***	1.028***	0.829***	0.833***
Pre		0.521***	0.535***	0.537***	0.582***
Condition			-0.180*	0.171	0.171
Time				0.010	0.010
Time*Condition				-0.018*	-0.017*
Pre*Condition					-0.092
Random effects					
Within-therapist variance	0.415***	0.325***	0.318***	0.308***	0.307***
Between-therapist variance	0.057	0.037	0.034	0.032	0.033
ICC	0.121	0.102	0.097	0.094	0.097
Model fit information					
Llg	-109.746	-95.092	-93.796	-92.125	-91.988
AIC	225.492	198.185	197.592	198.251	199.996
BIC	233.511	208.839	210.909	216.895	221.303

Note. All parameters are unstandardized correlation coefficients. TAU = treatment as usual condition; ROM = Routine Outcome Monitoring condition. Pre = pretreatment BASIS-32 scores (grand mean centered; M_{pre} = 1.44, SD_{pre}=0 .61); Condition = ROM (coded 1) or TAU (coded 0); Time = month of treatment start; ICC = Intraclass correlation coefficient; llg = loglikelihood; AIC = Akaike estimation; BIC = Bayesian estimation; ***p < .001; **p <. 01; *p < .05

Figure 2. The number of participants initiating treatment each month during the trial.

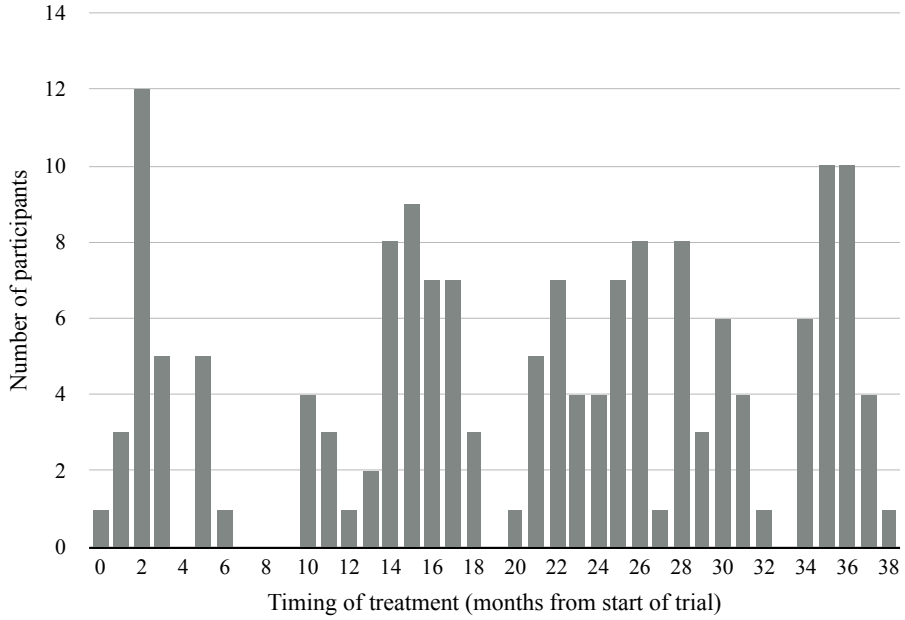
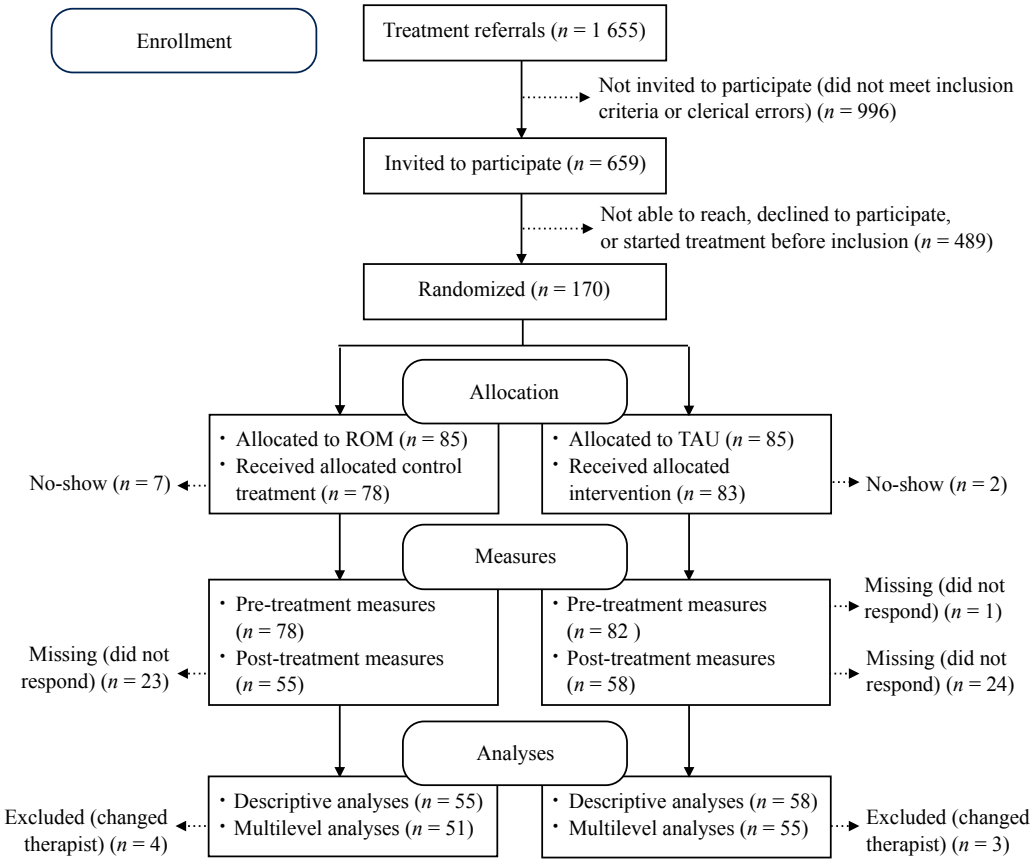


Figure 1. Participant flowchart.



Notes. TAU = treatment as usual condition; ROM = Routine Outcome Monitoring condition.