Original Article

CAMBRIDGE
UNIVERSITY PRESS

An evaluation of the efficacy of two add-on ecological momentary intervention modules for depression in a pragmatic randomized controlled trial (ZELF-i)

Jojanneke A. Bastiaansen[1,2] , Daan A. Ornée[1,2], Maaike Meurs[1] and Albertine J. Oldehinkel[1]

[1]Interdisciplinary Center Psychopathology and Emotion Regulation, Department of Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands and [2]Department of Education and Research, Friesland Mental Health Care Services, Leeuwarden, The Netherlands

Abstract

Background. Depression treatment might be enhanced by ecological momentary interventions (EMI) based on self-monitoring and person-specific feedback. This study is the first to examine the efficacy of two different EMI modules for depression in routine clinical practice.

Methods. Outpatients starting depression treatment at secondary mental health services ($N = 161$; $M_{\text{IDS–DEPRESSION}} = 35.9$, $S.D. = 10.7$; $M_{\text{AGE}} = 32.8$, $S.D. = 12.1$; 46% male) participated in a pragmatic randomized controlled trial with three arms. Two experimental groups engaged in 28 days of systematic self-monitoring (5 times per day), and received weekly feedback on either positive affect and activities (Do-module) or negative affect and thinking patterns (Think-module). The control group received no additional intervention. Participants completed questionnaires on depressive symptoms (primary outcome), social functioning, and empowerment before and after the intervention period, and at four measurements during a 6-month follow-up period.

Results. Of the 90 (out of 110) participants who completed the intervention, 86% would recommend it. However, the experimental groups did not show significantly more or faster changes over time than the control group in terms of depressive symptoms, social functioning, and empowerment. Furthermore, the trajectories of the two EMI modules were very similar.

Conclusions. We did not find statistical evidence that this type of EMI augments the efficacy of regular depression treatment, regardless of module content. We cannot rule out that EMIs have a positive impact on other domains or provide a more efficient way of delivering care. Nonetheless, EMI's promise of effectiveness has not materialized yet.

Introduction

Depression is the leading cause of disability worldwide and its prevalence is increasing (World Health Organization, 2017). Hence, top priorities in clinical mental health research are a better understanding of disease mechanisms and improved treatment efficacy (Elfeddali et al., 2014). Experience sampling and ecological momentary assessment (EMA) techniques (Larson & Csikszentmihalyi, 1983; Shiffman, Stone, & Hufford, 2008) have increased our general understanding of depression by providing novel insights into the daily emotional dynamics and physiological changes that accompany this condition (Myin-Germeys et al., 2009; Telford, McCarthy-Jones, Corcoran, & Rowse, 2012). More recently, researchers have suggested that these real-life self-assessments might also inform individual patients' clinical diagnosis and treatment (van Os, Delespaul, Wigman, Myin-Germeys, & Wichers, 2013a, b; Wichers, 2014). Moreover, self-monitoring with person-specific feedback has been put forward as a treatment in itself: it might reduce depressive symptoms by increasing self-awareness and inducing behavioral change (Myin-Germeys et al., 2018). Basic self-monitoring has been shown to improve emotional self-awareness, which enabled the recovery of depression (Kauer et al., 2012). EMA can provide an even more fine-grained film of the dynamics of depressive symptomatology, which can reveal previously implicit dysfunctional patterns and therefore provide new leads for behavioral change (Kramer et al., 2014). While studies have underscored the acceptability and feasibility of ecological momentary interventions (EMIs) with a self-monitoring component, there has been limited research on their efficacy (Colombo et al., 2019; Myin-Germeys, Klippel, Steinhart, & Reininghaus, 2016).

To date, only four EMIs have been evaluated in clinically depressed patients (Colombo et al., 2019) by studies with generally modest evidential strength. Two single-arm pilot studies

($n = 8$) of mobile applications incorporating symptom self-monitoring reported a decrease in depressive symptoms (Burns et al., 2011: Mobylize!; Mohr et al., 2015: MedLink). A pilot randomized controlled trial (RCT) that did include a control group ($n = 14$) in addition to an experimental group ($n = 14$) was not powered to statistically test group differences, but reported 'potentially meaningful improvements' in depressive symptoms among regular (but not casual) users[1] of their self-monitoring support platform (Burton et al., 2016: Help4Mood). Finally, a larger RCT ($n = 102$) reported that self-monitoring in addition to pharmacological treatment decreased depressive symptoms. These improvements were maintained over time in the group that received weekly feedback (Kramer et al., 2014: REsource MObilisation Device In Depression, REMOD-ID).

The EMI in the REMOD-ID study was based on a behavioral activation approach: the aim was to open up custom opportunities to increase the experience of positive affect (PA) by increasing personal insights in PA patterns and the context in which it is experienced (Kramer et al., 2014). While results were promising, the intervention had a substantial face-to-face component (6 weekly feedback sessions), which goes against the EMI principle of delivering psychological support in daily life (Colombo et al., 2019; Heron & Smyth, 2010) and could have driven treatment effects. Moreover, the intervention was evaluated in the relative absence of psychotherapy, a common part of depression treatment. Therefore, it remains unclear whether an EMI based on self-monitoring and person-specific feedback can add beneficial effects to regular depression treatment.

The first aim of the current study was to evaluate the efficacy of an EMI for depression in routine clinical practice. Early self-monitoring with personalized feedback might help patients obtain more insight in the processes involved in their depressive symptoms and day-to-day functioning, which might help them make the most of potential waitlist periods and commence treatment programs with a head start (Bastiaansen et al., 2018). Therefore, the patients who participated in this study started self-monitoring as soon as possible after clinical intake. Given that, for many patients, the essence of recovery is to rise above the presumed limitations associated with mental illness (Huber et al., 2011), we not only investigated the impact of an EMI on depressive symptomatology (primary outcome), but also on social functioning and feelings of empowerment; outcome domains that were also reported in complementary articles on the REMOD-ID study (Simons et al., 2015; Snippe et al.,).

The term EMI merely describes a method; the approach and content of self-monitored items and feedback vary from system to system (Burns et al., 2011; Burton et al., 2016; Kramer et al., 2014; Mohr et al., 2015), with unknown effects on efficacy. Hence, the second aim of the study was to examine the impact of EMI content on its efficacy. Participants who were randomized into one of the two intervention modules engaged in similar procedures, but the self-monitored items and feedback had a different focus: PA and activities in the 'Do'-module (reminiscent of the REMOD-ID study), and negative affect (NA) and thinking patterns in the 'Think'-module. These two EMI modules are both conceivably beneficial, as they link up with the two main angles of psychotherapy for depression: behavioral activation through positive reinforcement of activities and cognitive therapy to help individuals recognize and replace negative thinking patterns (Beck, Rush, Shaw, & Emery, 1987).

Our study is the first to examine the effects of – two different – EMI modules as an add-on to regular depression treatment. We hypothesized that the EMI groups would show more or faster positive changes over time compared to the control group (i.e. treatment-as-usual, TAU). We also looked into differences between the two EMI modules but did not have clear expectations of which one of the two would outperform the other based on the current literature.

Methods

Study design

The ZELF-i study was designed as a pragmatic RCT to allow evaluation of the intervention in real-life care facilities. The study protocol has been published elsewhere (Bastiaansen et al., 2018) and will be briefly explained below. The study was approved by the Medical Ethical Committee of the University Medical Center Groningen (UMCG, no. 2015/530). The trial has been registered prospectively in the Dutch Trial Register (Nederlands Trial Register, NTR5707, http://www.trialregister.nl) at 1 February 2016.

Participants

We recruited adult patients referred for depressive complaints to five general or specialized outpatient teams at four secondary mental health care organizations in the Netherlands. Specialized teams for affective disorders were asked to assess every new admission for eligibility, while general teams only assessed admissions listed as depression. Eligibility criteria were broad as to include a sample representative of clinical practice. Inclusion criteria were: (a) a clinical diagnosis of depression and primary indication for depression treatment by the mental health care professional (hereafter named: practitioner); (b) age between 18 and 65 years; and (c) written informed consent. Exclusion criteria were (based on practitioners' appraisals): (a) crisis intervention warranted; (b) presence of psychotic or manic symptoms; and (c) incapability of following research procedures due to inadequate Dutch language proficiency, significant auditory or visual impairments or mental retardation.

Randomization comprised a two-stage procedure. First, randomization was stratified per treatment location to account for clinical features that may influence outcomes: reported current psychotherapy (yes v. no) and antidepressant use in the 8 weeks prior to study entry (new/switch v. no/maintenance), similar to the REMOD-ID study. This stratification required the generation of 20 random allocation sequences (one for each of the four strata times five study locations) in stage two. For each sequence, block randomization was used to achieve balance in the allocation of participants to the study arms. That is, each of the three conditions was present twice (in random order) in a block of six. Participants were individually assigned to the control group or one of the two EMI groups (allocation ratio 1:1:1) by research assistants who did not have access to the random allocation sequences. The allocation was implemented by sequentially numbered sealed envelopes.

The sample size calculation was based on the primary outcome measure (depressive symptom severity) and indicated that 40 participants were needed in each of the three study groups (Bastiaansen et al., 2018, pp. 7–8). In the previous RCT (Kramer et al., 2014), the Inventory of Depressive Symptomatology Self Report (IDS-SR) scores for the experimental group showed an initial 3-point drop 8 weeks after baseline (Cohen's $f = 0.125$). With a sample size of 40 per group, an alpha of 0.05, an intraclass

correlation of 0.6 (between pre- and post-intervention measurements), and six measurements, the power to detect a group × time interaction of $f = 0.125$ was 97% (G×Power 3.1: Faul, Erdfelder, Lang, & Buchner, 2007). We considered smaller effects irrelevant and hence stopped recruitment when each study group included 40 'completers', that is, participants who completed the baseline, EMI (in case of the treatment groups), and post-EMA assessment. Data acquisition took place between May 2016 (first study intake) and March 2019 (last follow-up).

Intervention

Both EMA intervention modules comprised 28 consecutive days of systematic self-monitoring in combination with 4 weekly digital feedback reports and one face-to-face session to discuss the fourth and final feedback report with a research assistant. Eligible patients were enrolled in the study as soon as possible after clinical intake. Circumstances were otherwise kept as 'natural' as possible; regular treatment was not adapted and started upon availability, regardless of the study phase.

Self-assessments

Participants filled out brief questionnaires via a web application (RoQua, https://www.roqua.nl) on their smartphones. The measurements were set at five fixed moments during waking hours with an interval of 3 hours, programmed to optimally fit a participant's daily routine[2]. Each measurement comprised a momentary part, a module-specific retrospective part (past 3 hours) and a module-specific prospective (next 3 hours) part (for the full item list see Bastiaansen et al., 2018).

In both intervention modules, each measurement started with questions on momentary well-being, momentary affect (6 PA and 6 NA items), and momentary physical state, and ended with the question how much the participants were bothered by the measurement. In between, participants in the Do-module retrospectively recorded experienced pleasure, motivation, physical activity, busyness, time spent at home, in pleasant social contexts, and outdoors, and performed activities; and prospectively recorded anticipatory pleasure and motivation. Items deliberately focused[3] on positive contexts and activities to help participants monitor changes in their behavioral patterns, and ultimately increase their activity level, especially in pleasurable activities. Participants in the Think-module retrospectively recorded how much they focused on feelings, the amount of brooding, the occurrences of specific negative and positive events, and the presence of both negative and positive thoughts; and prospectively recorded worrying. Items were chosen to increase personal insights in daily events and participants' reactions to them with the ultimate goal of reducing NA.

In both modules, the morning measurement additionally included a question about sleep. Furthermore, the evening measurement also included four questions on how participants experienced the past day (retrospective well-being, coping, motivation, and mindfulness). Questions were mainly rated on visual analogue scales [usually ranging from not at all (0) to very much (100)], and in some cases on dichotomous scales (e.g. for activities and minor events).

Feedback reports

Standardized feedback reports (Online Supplementary Appendix A and https://osf.io/m6hvg/) were generated based on individual data and emailed to the participant after each week of EMA measurements with each successive report containing richer information. In line with behavioral activation, the Do-module reports comprised various graphs showing PA and activity patterns. In line with cognitive therapy, graphs in the Think-module focused on events, thinking patterns, and NA over time. The fourth report additionally included feedback on temporal relationships between sets of variables [e.g. PA and physical activity (Do-module[4]), or NA and rumination (Think-module)], but only for participants who filled out more than 75% of the measurements. This fourth and final feedback report was discussed with a research assistant in a face-to-face session in the week after the last EMA measurement (Online Supplementary Appendix A). Participants were encouraged to share the feedback report with their (future) therapist.

Measures

Demographic and clinical characteristics were queried at study intake (baseline). Participants completed online self-report questionnaires on depressive symptoms, psychosocial functioning, and empowerment at baseline (at the study site), in the week after the 28-day intervention period (post-EMA) and at four follow-up measurements 1, 2, 3, and 6 months post-EMA (at home). Participants in the treatment groups additionally completed an evaluation questionnaire (on site) at the post-EMA measurement. For the full list of questionnaires see Bastiaansen et al. (2018).

Depressive symptom severity

Change in depressive symptom severity was measured by the total score on the 30-item IDS-SR (Rush, Gullion, Basco, Jarrett, & Trivedi, 1996) across the six-time points. The IDS-SR includes all Diagnostic and Statistical Manual of Mental Disorders, fourth edition diagnostic criterion items for major depressive disorder, as well as commonly associated symptoms such as irritability. Each symptom item is scored on a scale from 0 to 3, with higher scores denoting greater symptom severity. The IDS-SR has good psychometric properties with high concurrent and internal validity and is sensitive to treatment change (Rush et al., 2003). In our study, Cronbach's alpha for the 30 items was 0.84 at baseline.

Social functioning

The Outcome Questionnaire-45 (OQ-45) is a 45-item self-report scale that measures subjective discomfort (SD), disturbance in interpersonal relations (IR) with partners, family and friends, and functioning in social roles (SR) such as work and school (Lambert et al., 1996). Each item is scored on a 5-point scale from never (0) to almost always (4). We administered – at each of the 6-time points – the Dutch version of the OQ, whose psychometric properties are adequate and similar to the original instrument (de Jong et al., 2007). In our analyses, we used the 11-item IR subscale (Cronbach's $\alpha = 0.69$ at baseline) and the 9-item SR subscale (Cronbach's $\alpha = 0.65$). Higher values on the IR and SR subscales indicate more disturbances in IR (range: 0–44) and SR functioning (range: 0–36), respectively.

Empowerment

The Netherlands Empowerment List [(NEL), Boevink, Kroon, & Giesen, 2008] is a 40-item self-rating scale to assess patient

empowerment, developed by the Dutch Trimbos Institute in collaboration with patients (for an English translation see van den Berg, van Amstel, Ottevanger, Gielissen, & Prins, 2013). Previous research has shown construct validity is satisfactory and internal consistency is high (Boevink et al., 2008). The NEL incorporates six dimensions: professional help, social support, own wisdom, sense of belonging, self-management, and community inclusion. Items are formulated in positive statements of strengths as perceived by the individual and are rated on 5-point scales ranging from 1 ('strongly disagree') to 5 ('strongly agree'). To prevent confounding by treatment status, we had to adjust our original plan by excluding the professional help subscale from the total empowerment score (36 items, range: 36–180, Cronbach's $\alpha = 0.87$ at baseline).

Start to treatment

We extracted information from the electronic patient records to determine care use and the time (in days) between study intake and the first psychotherapy session.

Statistical analysis

To reduce experimenter bias, analyses and data handling procedures were preregistered (https://osf.io/6kwre). The results of all preregistered analyses and (the rationale for) any deviations from the original preregistration are described in this article and related materials. The analysis code is openly available online (https://osf.io/m6hvg/).

The data had a hierarchical structure, with multiple assessments of the IDS-SR, OQ-45 and NEL being clustered within participants. We used R (R Core Team, 2018) and the lme4 (Bates, Mächler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017) packages to perform a multilevel regression analysis for each of the four outcome measures (IDS-SR, IR, SR, and NEL). Models included time (in months, not weeks as noted in the preregistration), group, and the two-way interaction between time and group as fixed effects; quadratic trends ($time^2$ and $group \times time^2$) were added to the model if they improved model fit. Models additionally included a random intercept and a random slope for time, which effectively allowed participants to vary in their experienced symptoms at baseline and in trajectories of change over time. We used full information maximum likelihood estimation, which can deal with data missing at random relatively well.

Our main analyses were based on the intention-to-treat principle: participants were compared within the groups to which they were initially randomized, independently of having received the allocated treatment, having dropped out of the study or having violated the initial protocol (for whatever reason). That is, participants were included in the main analyses regardless of the number of completed self-assessments, the number of feedback reports that were read, and whether the post-EMA feedback session was attended. In addition, we examined the efficacy of the add-on tool exclusively among participants who did not drop out of the intervention by means of a per-protocol analysis.

Results

Sample characteristics

Participant flow throughout the study is shown in Fig. 1. Approximately half of the eligible participants did not participate in the study; they either directly declined at clinical intake or were 'lost before study start' (i.e. they initially indicated interest but could not be reached by the research team or eventually did not attend or finish the study intake). The target sample size ($n = 120$ completers) was reached amply: 130 of the 161 patients randomized to one of the three study arms were completers. Ten controls did not complete the post-EMA measurement. One partial completer finished the 28-day intervention period but did not attend the post-EMA feedback session. Twenty participants dropped out of the intervention due to practical or time constraints (Do: 4, Think: 6), negative effects from completing the measurements (Do: 0, Think: 4), or for unknown reasons (Do: 3, Think: 3). There were no statistically significant differences (all $p > 0.31$) between participants who fully completed the intervention period and those who did not in baseline depressive symptoms, IR and SR functioning, and empowerment (Online Supplementary Appendix B). The three groups were very comparable in socio-demographic and clinical characteristics at baseline (Table 1). Regarding TAU, almost all participants (97%, $n = 155$) received a form of psychotherapy at one point during the study period. Participants mostly received (group or individual) cognitive behavioral therapy ($n = 108$) in combination with a wide array of other treatments. Sixty-seven participants (42%) used antidepressant medication at study start and 21 (13%) started, stopped or changed their medication during the intervention period.

Treatment adherence

Response compliance was high for the self-assessments (Table 2): after removal of dropouts, the average percentage of completed measurements was approximately 76% in both intervention modules. For half of the participants, response compliance was higher than 75%; they received additional feedback on temporal relationships in their final feedback report. The majority of participants read all weekly feedback reports and 92% intended to discuss the feedback with their therapist. However, at the 2-month follow-up (FU2), only 57% indicated to have actually done so.

Treatment evaluation

Detailed information on the patient-perceived feasibility and usability of the intervention can be found in Online Supplementary Appendix C. In brief, participants were positive about the usability of the web application. The feedback reports were rated fairly positively on comprehensibility and usefulness (≈ 60–70 out of 100) and contained the right amount of information according to the majority (86%). The face-to-face feedback session with a research assistant was perceived as very useful (≈ 80 out of 100). Of the 90 (out of 110) participants who completed the intervention, 86% would recommend it to others.

Treatment outcomes

All mixed model assumptions were satisfied (Online Supplementary Appendices D-G). Figure 2 displays the results of the multilevel regression analyses of the interaction between time and group for each of the four outcome measures.

Contrary to our primary hypothesis, neither of the intervention groups showed a significantly greater or faster decline over time in depression severity compared to the control group (Fig. 2a, Do v. Control: $B_{linear} = 0.1$, $t(538) = 0.1$, $p = 0.94$, and
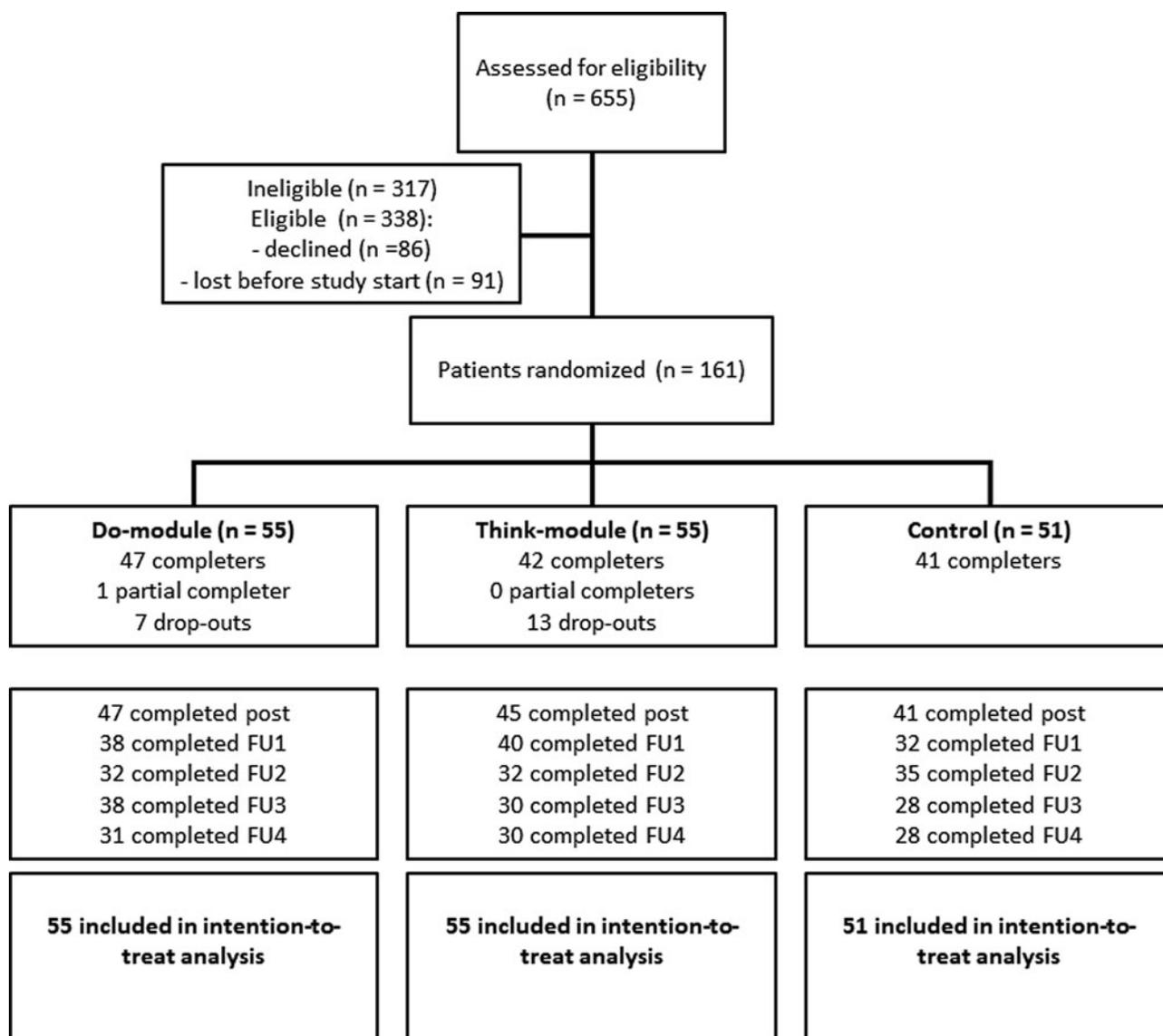
Fig. 1. Flow diagram of the ZELF-i study. *Note.* The diagram shows a multicenter randomized controlled trial with two treatment arms and one control arm. Post = assessment in the week after the 28-day EMA period (i.e. at the feedback session for the treatment groups). FU = follow-up assessment 1, 2, 3, and 6 months following the post-EMA assessment. Note that the number of completed measurements reflects depressive symptom assessments (IDS-SR); on a few occasions the depression questionnaire but not the social functioning (OQ-45) and/or empowerment (NEL) questionnaires were completed (max. $n = 3$ per FU, see Online Supplementary Appendices D-G).

$B_{quadratic} = 0.0$, $t(470) = 0.3$, $p = 0.79$; Think v. Control: $B_{linear} = 0.7$, $t(540) = 0.8$, $p = 0.44$, and $B_{quadratic} = -0.1$, $t(470) = -0.5$, $p = 0.59$). The Do-module and Think-module did not differ significantly from each other either ($B_{linear} = -0.7$, $t(539) = -0.7$, $p = 0.48$, and $B_{quadratic} = 0.1$, $t(471) = 0.8$, $p = 0.50$). Results were very similar for the per-protocol analysis (all interaction terms $p > 0.29$), which only included participants who completed the intervention (Do-module: $n = 48/55$, Think-module: $n = 42/55$). Overall, depression severity showed an early decline, which levelled off after FU1 (i.e. a combination of linear and quadratic trends). The average decline in depression severity from baseline to FU4 was 11 points on the IDS-SR with large between-person differences (S.D. = 14). Full analysis details, including the random effects (i.e. variation in individual effects), are presented in Online Supplementary Appendix D.

Full analysis details for disturbances in IR and SR functioning can be found in Online Supplementary Appendices E and F,

respectively. All groups showed a (modest) linear decrease in disturbances in IR (Fig. 2b) and in SR functioning (Fig. 2c) over time, and did not differ significantly from one another in the decline over time (IR scale, Do v. Control: $B_{linear} = 0.2$, $t(85) = 0.9$, $p = 0.36$; Think v. Control: $B_{linear} = -0.1$, $t(84) = -0.3$, $p = 0.73$; Do v. Think: $B_{linear} = 0.2$, $t(86) = 1.3$, $p = 0.20$; all per-protocol $p > 0.17$; SR scale, Do v. Control: $B_{linear} = -0.5$, $t(533) = -1.2$, $p = 0.25$, and $B_{quadratic} = 0.1$, $t(476) = 1.5$, $p = 0.14$; Think v. Control: $B_{linear} = -0.3$, $t(535) = -0.8$, $p = 0.41$, and $B_{quadratic} = 0.0$, $t(477) = 0.8$, $p = 0.42$; Do v. Think: $B_{linear} = -0.1$, $t(531) = -0.3$, $p = 0.74$, and $B_{quadratic} = 0.0$, $t(477) = 0.7$, $p = 0.51$; all per-protocol $p > 0.13$).

All groups showed an overall linear increase in empowerment over time, and did not differ significantly from one another (Fig. 2d, Do v. Control: $B_{linear} = -0.4$, $t(106) = -0.7$, $p = 0.46$; Think v. Control: $B_{linear} = -0.1$, $t(105) = -0.2$, $p = 0.83$; Do v. Think: $B_{linear} = -0.3$, $t(106) = -0.5$, $p = 0.60$; all per-protocol

Table 1. Baseline characteristics of study participants

	Total (n = 161)	Do-module (n = 55)	Think-module (n = 55)	Control (n = 51)
Age, yrs	32.8 ± 12.1	32.2 ± 11.6	33.6 ± 12.8	32.7 ± 12.0
Sex (M/F)	46/54%	49/51%	40/60%	49/51%
Educational level[a]				
Low	21%	27%	18%	18%
Middle	51%	49%	58%	45%
High	28%	24%	24%	37%
Daytime activity				
Study	24%	22%	22%	29%
Paid work	41%	40%	47%	35%
Household	4%	4%	4%	6%
No employment	30%	35%	27%	29%
Living with others	73%	80%	73%	65%
Previous depression treatment	48%	40%	53%	53%
Start to treatment, days[b]	19.7 ± 33.0	18.1 ± 30.7	23.8 ± 36.7	16.8 ± 31.4
Baseline antidepressant use[c]				
None	58%	62%	56%	57%
Maintenance	25%	22%	26%	27%
Change	17%	16%	18%	16%
Started	41%	33%	40%	50%
Stopped	41%	33%	40%	50%
Dosage change	48%	56%	50%	38%
Change in antidepressant use[d]	13%	16%	9%	14%
IDS-SR total score[e]	35.9 ± 10.7	35.7 ± 11.4	35.9 ± 10.5	36.2 ± 10.2
OQ-45 interpersonal relations	17.3 ± 5.5	16.5 ± 4.5	17.2 ± 5.2	18.3 ± 6.5
OQ-45 social roles	14.6 ± 4.9	13.7 ± 5.0	15.1 ± 4.4	14.9 ± 5.2
NEL empowerment[f]	110.2 ± 13.9	110.1 ± 15.4	110.5 ± 13.7	109.9 ± 12.6

Note. Numbers represent percentages or mean ± standard deviation. IDS-SR = Inventory of Depressive Symptomatology Self Report, OQ-45 = Outcome Questionnaire-45, NEL = NetherlandsEmpowerment List.

[a]Educational level – low: no/primary/low secondary, middle: high school/low vocational, high: higher vocational/university.

[b]Date of the first psychotherapy session at one of the study locations (as recorded in patients' files) minus the date of the study intake. We excluded the additionally preregistered self-report question on the start of psychotherapy, which proved difficult to answer (e.g. due to confusion with other appointments) and yielded many discrepancies with the patient files.

[c]Participants indicated whether they used antidepressants in the 8 weeks prior to study intake and whether the usage was stable (maintenance) or changed. In case of change, participants subsequently indicated which changes occurred (multiple responses possible).

[d]Change in antidepressant use between study intake and post-EMA assessment.

[e]Four participants (Do: 2, Think: 2, Control: 0) scored below the IDS-SR criterion for remission (*i.e.* 14: Meesters, Duijzer, Nolen, Schoevers, & Ruhé, 2016).

[f]The total empowerment score has been calculated excluding the professional help subscale (range: 36–180), which was not applicable at baseline for 24 participants in the Do-module, 25 participants in the Think-module, and 23 controls.

$p > 0.56$). Results were very similar if we used the empowerment measure – as was done in a previous study (Simons et al., 2015) – with 'imputed' scores for the professional help scale (although the Akaike information criterion favored a model with additional quadratic terms). Details for all empowerment analyses can be found in Online Supplementary Appendix G.

Post-hoc analyses

Detailed information on post hoc analyses can be found in Online Supplementary Appendix H. First, we explored depressive symptom trajectories in subgroups that were constructed based on compliance, and observed a more favorable course in the highly compliant group (≥75%) compared to the less compliant group.

Given the intermediate trajectory of the control group (who did not complete any repeated self-assessments), compliance is more likely to be a marker for a favorable course than its cause. Second, we explored whether the EMI might have added value in the absence of TAU: we observed that although participants who engaged in the EMI while waiting for psychotherapy showed early symptom declines, similar declines were seen in controls waiting for psychotherapy. Formal testing was impossible because subgroups were small and the waitlist condition was not 'clean': before the start of psychotherapy patients turned out to often have had other appointments (e.g. diagnostic testing, medication consults). Third, we merged the figure comprising depressive outcome data of the experimental group and control group of the previous RCT (Kramer et al., 2014) with our outcome data.

Table 2. Treatment adherence

	Total	n[a]	Do-module	n[a]	Think-module	n[a]
% Diaries completed						
Total sample	66.0 ± 29.0	110	70.5 ± 24.9	55	61.5 ± 32.1	55
Completers	76.7 ± 18.6	110	76.9 ± 18.4	48	76.4 ± 19.2	42
Dropouts	18.0 ± 14.8	110	26.6 ± 19.1	7	13.4 ± 9.9	13
Compliance >75%	49%	110	56%	55	42%	55
Diary duration, sec	119 [88;167]	10 314	115 [87;162]	5505	124 [90;173]	4809
Feedback reports						
Read week 1	94%	90	91%	47	100%	43
Read week 2	92%	90	87%	47	98%	43
Read week 3	71%	90	70%	47	72%	43
Read week 4	78%	90	77%	47	79%	43
Intention to discuss in therapy (post-EMA)	92%	90	94%	47	91%	43
Discussed in therapy (FU2)[b]	57%	60	58%	31	55%	29

Note. Numbers represent percentage, mean ± standard deviation, or median [25th;75th percentile]. FU2 = the follow-up assessment 2 months after the post-EMA assessment.
[a]The sample size differs per variable due to missing data or different measurement levels (i.e. diary duration is based on the number of valid measurements rather than the number of participants).
[b]Reasons why participants did not discuss the feedback report with their therapist were: treatment did not start yet ($n = 8$), report not useful ($n = 5$), did not know how to discuss it ($n = 4$), did not finish the intervention ($n = 2$), or other ($n = 7$, *e.g.* forgotten, never came up, did not get around to it yet).

Upon reviewer request, we also reran our intention-to-treat analyses including covariates (educational level, type of TAU, etc.) to study the impact on efficacy estimates. Estimates were essentially the same as those of the original model, with none of the interaction terms reaching significance. This indicates that the original comparison of treatment groups was reasonable. Furthermore, we added separate comparisons between baseline and each of the five follow-up measurements (post-EMA, FU1-4), in which each measurement (e.g. FU1) was regressed onto the baseline measure with the group as an added predictor. Group differences were small for each of the individual comparisons (for all outcome measures); none of the group comparisons reached an uncorrected significance level of 0.05, let alone after Bonferroni adjustment ($p < 0.01$).

Discussion

Overall, the study participants showed significant improvements over time in depressive symptoms, social functioning, and empowerment. However, the EMI groups did not show more or faster changes over time than the control group. Furthermore, the trajectories of the two EMI modules were very similar. Hence, we did not find statistical evidence that EMIs based on self-monitoring and person-specific feedback could augment the efficacy of regular depression treatment.

Our results seem to stand in sharp contrast with the only previous RCT (Kramer et al., 2014: REMOD-ID), which concluded that an EMI could be an effective therapeutic tool for depression. However, differential outcomes between these studies appear to relate less to the response in the experimental groups – the symptom decline in these groups was rather comparable across studies (Online Supplementary Appendix H) - than to the symptom trajectories of the control groups: a flat line for depressive patients receiving pharmacotherapy only (REMOD-ID) and a decline for those receiving TAU including psychotherapy (ZELF-i). The

EMI groups of neither study seemed to outperform the latter control group.

More specifically, the EMI group from the REMOD-ID study that did not receive psychotherapy showed changes comparable to our control group, which received no EMI but psychotherapy. Together, these studies suggest that EMIs activate similar mechanisms as psychotherapy. This is well conceivable given that the EMIs in both studies were based on common behavioral or cognitive strategies that have been found effective in the treatment of depression (Cuijpers, Karyotaki, de Wit, & Ebert, 2019). Although cognitive and behavioral therapies are more comprehensive, EMIs might be seen as a sophisticated extension of the self-assessment homework that psychotherapy usually entails already: paper-and-pencil activity monitoring and ABC (antecedent-beliefs-consequences) worksheets. Whereas targeted depression mechanisms may be similar, the EMI approach specifically aims to mobilize patients as active agents in their recovery process (Myin-Germeys et al., 2018; van Os et al., 2013*a*, *b*), and hence might be expected to enhance empowerment. However, we did not find evidence for an EMI-specific empowerment increase. In the REMOD-ID study, empowerment increases in the experimental group were not significantly different from the control group either ($p = 0.061$; Simons et al., 2015).

If the activated mechanisms of EMI and psychotherapy are similar, EMIs might primarily be effective when interventions involving face-to-face sessions are not – yet – available. The REMOD-ID findings are not conclusive in this respect, despite the comparison with passive pharmacotherapy, because its EMI involved repeated face-to-face feedback sessions itself. Our study could not provide a definite answer on whether the EMI had added value in the absence of TAU because subgroups that had to wait for psychotherapy oftentimes did have intermediate consultations. Whereas additional studies in groups with no actual access to psychotherapy could be useful in this matter, we believe a more fruitful endeavor would be to examine whether EMIs could partly substitute standard psychological treatment without
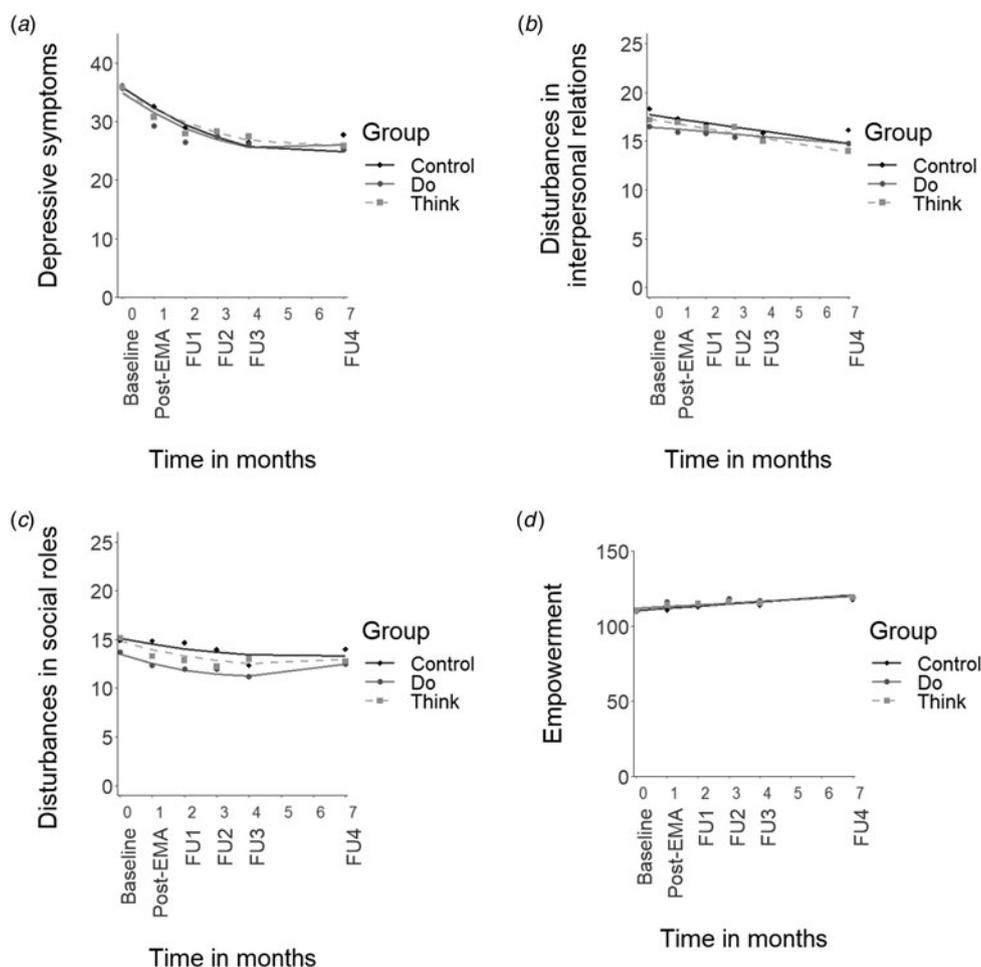
Fig. 2. Mean scores for the outcome measures and predicted lines plotted across time. *Note.* (a) depressive symptom severity (IDS-SR total score), (b) disturbances in interpersonal relations (OQ-45 IR scale), (c) disturbances in social role functioning (OQ-45 SR scale), (d) empowerment (NEL total score without the professional help scale). Post-EMA = assessment after the 28-day intervention period. FU = follow-up assessment 1, 2, 3, and 6 months following the post-EMA assessment. Note that the time between the baseline and post-EMA measurement spanned approximately 5–6 weeks rather than 1 month (i.e. post-EMA measurements were typically planned in the week after the 28-day intervention period).

sacrificing efficacy. The only non-inferiority trial in depressed patients to date tentatively suggests that a blended treatment – including four face-to-face sessions and a self-monitoring smartphone application – could possibly treat nearly twice as many depressed patients compared to a full behavioral activation treatment, with comparable results (Ly et al., 2015). Thus, perhaps the promise of EMI might not be increased *efficacy*, but a more *efficient* way of delivering care, which could reach many more patients in need.

Although we did not find evidence for an EMI effect on depressive symptoms, empowerment, or social functioning, we cannot exclude the possibility that our EMI modules had an impact on other domains. In fact, the high percentage of study participants who reported that they would recommend the intervention seems to suggest that they did experience some utility from it (although we should note that such a subjective measure comes with a risk of overestimation due to, for instance, agreement bias: Chang, Gillespie, & Shaverdian, 2019). EMIs might, for instance, help patients acquire better self-insight. A recent study showed that the experimental groups of the REMOD-ID study improved in emotion differentiation (Widdershoven et al., 2019). Furthermore, an intervention study in adolescents with emotional problems, which included a self-report measure of

emotional self-awareness, found that self-monitoring had a direct effect on emotional self-awareness (Kauer et al., 2012: mobiletype). Moreover, changes in depressive symptoms were mediated by increases in emotional self-awareness in the intervention group. Thus, although an EMI may not lead to better mood *per se*, it might advance self-insight, which could be a first step on the road to recovery. Future research needs to evaluate whether we are studying the right outcome domains. Qualitative work on the impact of EMIs from patients' perspectives might be a good starting point.

Strong suits of this study are the rigorous study design (RCT) comprising two highly comparable intervention modules, the inclusion of clinical as well as functional outcome measures across multiple time points, the good treatment adherence, and the naturalistic setting, which allows the generalization of results to regular clinical practice. A drawback of the naturalistic setting is the resulting heterogeneity of the TAU condition (in both timing and content) that may have added noise and leaves open the possibility that the EMI is effective under certain conditions. In addition, given that therapist guidance might bolster the effectiveness of smartphone interventions (Linardon, Cuijpers, Carlbring, Messer, & Fuller-Tyszkiewicz, 2019), the EMI might have had a stronger effect if it had been more integrated with

psychotherapy: about half of the patients who received the EMI indicated they did not discuss their feedback reports with their therapist. Furthermore, our final sample comprised only patients who were willing and able to participate in a research study beside starting their regular treatment. It is unlikely that this inevitable selection influenced the main outcomes of our trial. If anything, one would expect an overestimation – not absence – of a treatment effect. Finally, it is important to highlight that this study regarded only one type of EMI; other clinical applications are conceivable and might be more effective.

To conclude, we did not find statistical evidence that the EMI impacted clinical or functional outcomes beyond the effects of TAU, regardless of module content. This does not rule out that EMIs could have a positive effect on other domains or provide a more *efficient* way of delivering care. However, EMIs promise of *effectiveness* has not materialized yet.

Notes

1 Participants were considered 'regular' users if they used Help4Mood 10 times or more during a four-week period and 'casual' users if they used the platform on 3-7 days (Burton et al., 2016).

2 Each participant's measurement schedule comprised five assessments spaced three hours apart. This schedule was set (in steps of 15 minutes) in consultation with the participant to accommodate usual bedtime, wake up time, and recurrent appointments (e.g. classes or meetings starting fifteen minutes past the hour). Due to technical restrictions, the last questionnaire had to be completed before midnight and was sent no later than 11.30 pm (30-minute response window), which also restricted the final timing of the first measurement to 11.30 am. In reality, the median for the first measurement was 10 am (IQR = 9.15 – 10.45 am).

3 Note that the scope of the Do-module was comparable to the REMOD-ID study, but items were not exactly alike. The REMOD-ID study (Kramer et al.,

2014), for instance, measured current affect with 4 PA and 6 NA items, and assessed current context and activities with slightly different questions.

4 Note again that the scope of the Do-module is comparable to the REMOD-ID study, but feedback reports were not exactly alike. Feedback in the REMOD-ID study, for instance, only comprised descriptive statistics and not the results of individual models.

References

Bastiaansen, J. A., Meurs, M., Stelwagen, R., Wunderink, L., Schoevers, R. A., Wichers, M., & Oldehinkel, A. J. (2018). Self-monitoring and personalized feedback based on the experiencing sampling method as a tool to boost depression treatment: A protocol of a pragmatic randomized controlled trial (ZELF-i). *BMC Psychiatry*, *18*(1), 276. doi:10.1186/s12888-018-1847-z.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01.

Beck, A. T., Rush, A. J., Shaw, B., & Emery, G. (1987). *Cognitive therapy of depression*. New York, NY, US: Guilford Press. Retrieved from https://www.guilford.com/books/Cognitive-Therapy-of-Depression/Beck-Rush-Shaw-Emery/9780898629194.

Boevink, W., Kroon, H., & Giesen, F. (2008). *De Nederlandse empowerment vragenlijst*. Utrecht: Trimbos Instituut.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, *13*(3), e55. doi:10.2196/jmir.1838.

Burton, C., Szentagotai Tatar, A., McKinstry, B., Matheson, C., Matu, S., Moldovan, R., … for the Help4Mood Consortium. (2016). Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. *Journal of Telemedicine and Telecare*, *22*(6), 348–355. doi:10.1177/1357633X15609793.

Chang, E. M., Gillespie, E. F., & Shaverdian, N. (2019). Truthfulness in patient-reported outcomes: Factors affecting patients' responses and impact on data quality. *Patient Related Outcome Measures, Volume*, *10*, 171–186. doi:10.2147/PROM.S178344.

Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., García-Palacios, A., … Botella, C. (2019). Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: A systematic review. *Journal of Clinical Medicine*, *8*(4), 465. doi: 10.3390/jcm8040465.

Cuijpers, P., Karyotaki, E., de Wit, L., & Ebert, D. D. (2019). The effects of fifteen evidence-supported therapies for adult depression: A meta-analytic review. *Psychotherapy Research*, *30*(3), 279–293. doi:10.1080/10503307.2019.1649732.

de Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy*, *14*(4), 288–301. doi:10.1002/cpp.529.

Elfeddali, I., van der Feltz-Cornelis, C. M., van Os, J., Knappe, S., Vieta, E., Wittchen, H.-U., … Haro, J. M. (2014). Horizon 2020 priorities in clinical mental health research: Results of a consensus-based ROAMER expert survey. *International Journal of Environmental Research and Public Health*, *11*(10), 10915–10939. doi:10.3390/ijerph111010915.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*(1), 1–39. doi:10.1348/135910709X466063.

Huber, M., Knottnerus, J. A., Green, L., Horst, H. V. D., Jadad, A. R., Kromhout, D., … Smid, H. (2011). How should we define health? *BMJ*, *343*, d4163. doi:10.1136/bmj.d4163.

Kauer, S. D., Reid, S. C., Crooke, A. H. D., Khor, A., Hearps, S. J. C., Jorm, A. F., … Patton, G. (2012). Self-monitoring using mobile phones in the early stages of adolescent depression: Randomized controlled trial. *Journal of Medical Internet Research*, *14*(3), e67. doi:10.2196/jmir.1858.

Kramer, I., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., … Wichers, M. (2014). A therapeutic application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry*, 13(1), 68–77. doi:10.1002/wps.20090.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(1), 1–26. doi:10.18637/jss.v082.i13.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, 3(4), 249–258. doi:10.1002/(SICI)1099-0879(199612)3:4 < 249::AID-CPP106>3.0.CO;2-S.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *New directions for methodology of social and behavioral sciences* (Vol. 15, pp. 41–56). San Francisco: Jossey-Bass.

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials. *World Psychiatry*, 18(3), 325–336. doi:10.1002/wps.20673.

Ly, K. H., Topooco, N., Cederlund, H., Wallin, A., Bergström, J., Molander, O., … Andersson, G. (2015). Smartphone-supported versus full behavioural activation for depression: A randomised controlled trial. *PLoS One*, 10(5), e0126559. doi:10.1371/journal.pone.0126559.

Mohr, D. C., Stiles-Shields, C., Brenner, C., Palac, H., Montague, E., Kaiser, S. M., … Duffecy, J. (2015). MedLink: A mobile intervention to address failure points in the treatment of depression in general medicine. International Conference on Pervasive Computing Technologies for Healthcare, 2015, 100–107. doi: 10.4108/icst.pervasivehealth.2015.259042.

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry*, 17(2), 123–132. doi:10.1002/wps.20513.

Myin-Germeys, I., Klippel, A., Steinhart, H., & Reininghaus, U. (2016). Ecological momentary interventions in psychiatry. *Current Opinion in Psychiatry*, 29(4), 258–263. doi:10.1097/YCO.0000000000000255.

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: Opening the black box of daily life. *Psychological Medicine*, 39(9), 1533–1547. doi:10.1017/S0033291708004947.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The inventory of depressive symptomatology (IDS): Psychometric properties. *Psychological Medicine*, 26(3), 477–486. doi:10.1017/s0033291700035558.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., … Keller, M. B. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5), 573–583. doi:10.1016/S0006-3223(02)01866-8.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4(1), 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415.

Simons, C. J. P., Hartmann, J. A., Kramer, I., Menne-Lothmann, C., Höhn, P., van Bemmel, A. L., … Wichers, M. (2015). Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. *European Psychiatry*, 30(8), 900–906. doi:10.1016/j.eurpsy.2015.09.004.

Snippe, E., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Kramer, I., Booij, S. H., … Wichers, M. (2015). Change in daily life behaviors and depression: Within-person and between-person associations. *Health Psychology*, 35(5), 433–441. doi:10.1037/hea0000312.

Telford, C., McCarthy-Jones, S., Corcoran, R., & Rowse, G. (2012). Experience sampling methodology studies of depression: The state of the art. *Psychological Medicine*, 42(6), 1119–1129. doi:10.1017/S0033291711002200.

van den Berg, S. W., van Amstel, F. K. P., Ottevanger, P. B., Gielissen, M. F. M., & Prins, J. B. (2013). The cancer empowerment questionnaire: Psychological empowerment in breast cancer survivors. *Journal of Psychosocial Oncology*, 31(5), 565–583. doi:10.1080/07347332.2013.825361.

van Os, J., Delespaul, P., Wigman, J., Myin-Germeys, I., & Wichers, M. (2013a). Beyond DSM and ICD: Introducing 'precision diagnosis' for psychiatry using momentary assessment technology. *World Psychiatry*, 12(2), 113–117. doi:10.1002/wps.20046.

Van Os, J., Delespaul, P., Wigman, J., Myin-Germeys, I., & Wichers, M. (2013b). Psychiatry beyond labels: Introducing contextual precision diagnosis across stages of psychopathology. *Psychological Medicine*, 43(7), 1563–1567. doi:10.1017/S0033291713000937.

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, 44(07), 1349–1360. doi:10.1017/S0033291713001979.

Widdershoven, R. L. A., Wichers, M., Kuppens, P., Hartmann, J. A., Menne-Lothmann, C., Simons, C. J. P., & Bastiaansen, J. A. (2019). Effect of self-monitoring through experience sampling on emotion differentiation in depression. *Journal of Affective Disorders*, 244, 71–77. doi:10.1016/j.jad.2018.10.092.

World Health Organization (2017). *Depression and other common mental disorders: Global health estimates*. Geneva: World Health Organization. Retrieved from World Health Organization website: https://apps.who.int/iris/handle/10665/254610.