Opportunities and Challenges of Ecological Momentary Assessment for Research on Psychological Interventions in Depression

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie

an der Ludwig-Maximilians-Universität

München

Fakultät für Psychologie & Pädagogik

Lehrstuhl für Klinische Psychologie & Psychotherapie

vorgelegt von

Jeanette Tamm

aus

Bad Soden

München, 2025

Erstgutachter: Prof. Dr. Thomas Ehring

Zweitgutachterin: PD Dr. Belinda Weber

Tag der mündlichen Prüfung: 12.02.2025

Opportunities and Challenges of Ecological Momentary Assessment for Research on Psychological Interventions in Depression

Danksagung

Ohne die Unterstützung zahlreicher Personen, die meine Promotion begleitet haben, wäre die Realisierung dieser Dissertation in ihrer Form nicht möglich gewesen. Mein besonderer Dank gilt folgenden Personen:

An erster Stelle danke ich Herrn Prof. Dr. Thomas Ehring für sein Vertrauen in meine Kompetenzen, Ideen und Forschungsinteressen. Ich danke ihm für seine exzellente fachliche und menschliche Unterstützung, inklusive seines Verständnisses und seinen Ermutigungen in schwierigen Phasen. Gleichermaßen danke ich Herrn Dr. Johannes Kopf-Beck für die engmaschige Betreuung und Supervision meiner Promotion auf fachlicher und menschlicher Ebene. Darüber hinaus danke ich ihm für den besonderen Zusammenhalt der OPTIMA Arbeitsgruppe, z.B. in Zeiten der Corona Pandemie und das bis heute stets gebührende Feiern aller Meilensteine und Erfolge des Studienteams. Darüber hinaus danke ich Herrn Dr. Keisuke Takano für sein exzellentes fachliches Feedback, dessen Expertise ich sehr bewundere. Ich danke zudem Frau Dr. Tabea Rosenkranz für ihre fachliche Unterstützung am Anfang meiner Promotion.

Ich danke dem gesamten OPTIMA Studienteam für die jahrelange Zusammenarbeit, die in besonderem Maße von gegenseitiger Unterstützung geprägt war. Ein spezieller Dank gilt dabei Leah Just und Sarah Kirchler, die meine Promotion besonders unterstützt haben sowie Celina Müller, Zoe Spock und Gabi Kohl, für den Zusammenhalt in der Studie und darüber hinaus. Ich danke zudem allen Studienteilnehmer*innen für ihre Einsatzbereitschaft und ihr Vertrauen in unsere Forschung. Desweiteren möchte ich meinen Dank an Herrn Dr. Victor Spoormaker aussprechen, von dem ich im Anschluss viel gelernt habe. In dem Zusammenhang danke ich allen Kolleg*innen aus seiner Forschungsgruppe, für die großartige Zusammenarbeit in den letzten 2,5 Jahren, die meine Promotion ebenfalls geprägt hat. Ein besonderer Dank gilt dabei Frau Dr. Julia Fietz, Daniel Pfahl und Andy Brendler.

Auf der privaten Seite möchte ich mich in erster Linie bei meiner Familie bedanken, auf deren Unterstützung ich mich immer verlassen kann. Ich danke euch für euer Urvertrauen in meine Fähigkeiten, für eure Motivationsschübe, eure warmen Worte und euer Mitfiebern bei allen Meilensteinen in meinem Leben, inklusive meiner Promotion. Daneben möchte ich mich bei meinen Freundinnen und Freunden bedanken, die immer an meiner Seite stehen, ein offenes Ohr für mich haben und alle meine Erfolge mit mir feiern. Ein besonderer Dank gilt

dabei Frau Dr. Paulina Dembinski, die mit mir die Höhen und Tiefen des Promovierens gemeinsam durchlebt hat und daher alle Probleme und Sorgen besonders gut nachvollziehen konnte. Ich danke dir für all deine Motivationsschübe, das gemeinsame Lachen (und Weinen) sowie deine Ratschläge, die ich in privater und beruflicher Hinsicht immer schätze.
Schließlich möchte ich meinem Partner für seine Unterstützung danken. Vielen Dank für deinen Halt in den schwierigen Phasen, deine unterstützenden Worte, das gemeinsame Genießen unserer Erfolge sowie deine stets liebevollen Ratschläge.

Author Contributions

This is a cumulative doctoral thesis presenting the results of an empirical quantitative study, which was conducted at the Max Planck Institute of Psychiatry in Munich. This thesis includes three scientific papers about the study, which were supervised by Prof. Dr. Thomas Ehring, Dr. Johannes Kopf-Beck, Dr. Keisuke Takano and Dr. Tabea Rosenkranz. The scientific papers have been published or submitted for publication in international peer-reviewed journals. The authors' contributions to these papers were as follows:

Study I: Tamm, J., Takano, K., Just, L., Ehring, T., Rosenkranz, T. & Kopf-Beck, J. (2024) Ecological Momentary Assessment versus Weekly Questionnaire Assessment of Change in Depression. Depression and Anxiety, 2024(1), 9191823. https://doi.org/10.1155/2024/9191823.

Jeanette Tamm conceptualized the study, administered the project and data collection, curated the data, designed the methodology, performed the data analysis, created the visualizations and drafted the manuscript. The co-author JKB administered the umbrella study (OPTIMA). JKB, KT and TE supervised the conceptualization, methodology and analysis of the study, and reviewed the manuscript. LJ supported the collection and curation of the data, as well as the language editing of the manuscript. TR contributed to the supervision of the study conceptualization and methodology.

Study II: Tamm, J., Takano, K., Just, L., Ehring, T., Rosenkranz, T., OPTIMA study group & Kopf-Beck, J. (under review). Early Improvement predicts Treatment response in Depression: An Ecological Momentary Assessment Study. Manuscript submitted to *Behavior Therapy*.

Jeanette Tamm conceptualized the study, administered the project and data collection, curated the data, designed the methodology, performed the data analysis, created the visualizations and drafted the manuscript. The co-author JKB administered the umbrella study (OPTIMA). JKB, KT and TE supervised the conceptualization, methodology and analysis of the study, and reviewed the manuscript. LJ supported the collection and curation of the data, as well as the language editing of the manuscript. TR contributed to the supervision of the study conceptualization.

Study III: Kirchler, S. V., Müller, C. L., Spock, Z., Ehring, T., Kopf-Beck, J.* & Tamm, J.* (under review). The Role of Concreteness in Repetitive Negative Thinking: Temporal

Dynamics and the Predictive Value for Depression Throughout Psychological Treatment. Manuscript submitted to *Behaviour Research and Therapy*.

** The last two authors contributed equally to this work.*

Jeanette Tamm conceptualized the overall research question, administered the data collection, curated part of the data, supervised the methodology and data analysis, validated the analysis and results and reviewed the manuscript. The first author, SVK, conceptualized the methodology, administered the data curation and analysis, and drafted the manuscript. The co-author JKB administered the umbrella study (OPTIMA). JKB and TE supported the supervision of the conceptualization, methodology and analysis of the study, and reviewed the manuscript. CLM and ZS contributed to the data collection and curated part of the data. In addition, CLM supported the data analysis, and ZS the language-editing.

Abstract

Ecological Momentary Assessment (EMA) has gained increasing popularity in depression research due to its ability to capture symptoms in real time and its potential to mitigate recall bias present in retrospective clinical assessments. In our EMA substudy, conducted within a large randomized controlled trial comparing three psychotherapeutic interventions - cognitive behavioral therapy (CBT), schema therapy (ST), and individual supportive therapy (IST) - we examined the opportunities and challenges of EMA in supporting psychotherapy research and practice in depression. Over the course of seven weeks of psychotherapy, 106 moderately to severely depressed patients provided momentary self-reports of depressive symptoms and repetitive negative thinking (RNT) three times daily. In addition, a comprehensive test battery, including weekly questionnaire assessments (WQA) of depressive symptoms and RNT, and clinical interviews of global functioning, was assessed before and at the end of the intervention. RNT is a transdiagnostic cognitive process that plays an important role in the development and maintanence of depression. Defined as repetitive, intrusive, relatively uncontrollable, and of negative content, RNT is an umbrella term for rumination and worry. The collected data was analyzed in three different studies:

Study I compared the results of EMA and WQA in terms of measuring changes in depressive symptoms and RNT. We found that EMA was more sensitive to detecting between-group differences in intervention effects. Specifically, it revealed a superior reduction of RNT in the ST group compared to CBT and IST, which was not detected by WQA. The higher sensitivity of EMA for intervention effects may stem from a higher measurement reliability due to its real-time assessments, which avoid recall bias inherent in retrospective questionnaires. In contrast, WQA proved more effective in predicting changes in clinician-rated global functioning, potentially due to the common retrospective nature of the two measures. These findings highlight the complementary strengths of EMA, WQA, and clinical interviews and suggest that integrating these methods into clinical assessments could accelerate the comparison of intervention effects in clinical trials by improving measurement reliability.

Study II focused on predicting treatment response (versus non-response) based on early improvements in depressive symptoms as assessed by EMA versus WQA. Our analyses showed that early improvements assessed by either method significantly predicted treatment response within three weeks of treatment initiation. However, WQA provided a clearer

pattern of the optimal predictive change rate, indicating that a 10% symptom improvement at four weeks resulted in a true negative rate of 22% compared with a false negative rate of 0%. EMA provided comparable predictive power but lacked clarity in its pattern of an optimal predictive change rate. These findings demonstrate the potential of WQA and EMA for early treatment prediction, while suggesting that the clarity of the prediction pattern may depend on the measure of treatment response, which in our study was equivalent to the WQA predictor (both were operationalized with the BDI-II).

Study III used EMA to explore the temporal relationships between momentary levels of depressive symptoms and concreteness levels of RNT, as well as their changes over the intervention course. Depressed patients tend to ruminate and worry in a less concrete manner than healthy individuals, i.e., their RNT is more unclear, aggregated, cross-situational and less solution-oriented. Our study showed that RNT concreteness explains additional variance in momentary levels of depressive symptoms that is not explained by the process of RNT itself. Notably, changes in RNT concreteness over the course of therapy interacted with patients' improvement in depression severity: patients who improved more than average showed a slight increase in concreteness, while those who improved less showed a decrease. In addition, higher levels of momentary depressive symptoms predicted subsequent decreases in momentary levels of concreteness, but not vice versa. These findings suggest that future studies should examine the long-term dynamics between RNT concreteness and depression.

Based on these investigations, several strategies for refining EMA approaches in future studies to improve data quality and patient adherence are discussed. In all three studies, EMA’s ability to capture real-time fluctuations in depressive symptoms and RNT provided new insights into the assessment, treatment and understanding of depression. In addition, the combination of EMA with emerging technologies, such as passive data tracking and AI-based text analyses to automate complex rating procedures such as RNT concreteness ratings, offers significant potential for advancing EMA approaches. Besides, providing continuous personalized feedback on symptom progression, delivering just-in-time recommendations, and optimizing treatment module allocation based on EMA data are promising strategies for developing personalized, potentially more effective treatments.

The temporal dynamics between depressive symptoms measurable with EMA support a growing shift from traditional latent-disease models to a network perspective on mental disorders, in which transdiagnostic factors like RNT and global functioning gain an increased role. Nevertheless, the rapid adoption of disruptive technologies like EMA and AI

underscores the need for careful investigation of their opportunities and challenges in psychotherapy research and practice for depression.

Table of Contents

Abstract VIII

1. General Introduction 13

2. Study I: *Ecological Momentary Assessment versus Weekly Questionnaire Assessment of Change in Depresion* **..... 31**

3. Study II: *Early Improvement predicts Treatment Response in Depression: An Ecological Momentary Assessment Study* **..... 57**

4. Study III: *The Role of Concreteness in Repetitive Negative Thinking: Temporal Dynamics and the Predictive Value for Depression Throughout Psychological Treatment* **..... 81**

5. General Discussion 107

Zusammenfassung 131

Abbreviations 140

References 141

List of Figures 170

List of Tables 171

Appendix A: Supplementary Material Study I 173

Appendix B: Supplementary Material Study II 181

Appendix C: Supplementary Material Study III 189

1. General Introduction

Over the past century, our understanding of depression has expanded significantly, fueled by advancements in psychological theories and interventions, the development of antidepressant medications, and breakthroughs in neuroimaging. Now, as we look to the future, the global digitalization and an increasing focus on the individual hold significant potential to improve our understanding of depression and open new avenues to develop more precise assessments and personalized treatments.

Depression – a highly prevalent and complex mental disorder

Depression is the second most prevalent mental health disorder worldwide, and the number of people affected is reaching new highs after Covid-19 (Santomauro et al., 2021). According to latest estimates, approximately 300 million people, or more than 5% of the global adult population, are now living with depression (GBD 2019 Mental Disorders Collaborators, 2022; Arias-de la Torre et al., 2021), and up to 21% will experience depression at some point in their lives (Gutiérrez-Rojas, 2020). The World Health Organization (2017) rates depression the leading cause of ill health and disability worldwide, causing enormous patient burden and economic costs each year (Arias et al., 2022; König et al., 2020). Despite the existence of effective treatments for depression, including antidepressants and psychotherapy (Cuijpers, Oud et al., 2021; Cuijpers, Miguel et al., 2023), nearly 50% of patients remain untreated (Mekonen et al., 2021), and of those receiving first-line treatment, only about 40% achieve a response and 30% achieve remission (Cuijpers, Karyotaki et al., 2021). It is therefore a global societal challenge to understand the mechanisms underlying depression, to improve the distribution of mental health care (WHO, 2017), and to develop more effective treatments.

In addition to its high prevalence, depression is a highly complex mental disorder. According to contemporary classification systems such as the ICD-11, to fulfill the diagnosis of major depression, one must experience low mood and/or decreased interest in activities, which constitute the core symptoms of depression, in conjunction with eight optional additional symptoms such as concentration problems, changes in appetite and/or sleep, or psychomotor inhibition or agitation (WHO, 2022). This means that depression can theoretically manifest in several thousand possible unique symptom profiles (Fried & Nesse, 2015), and the high prevalence of comorbid mental disorders, most commonly anxiety disorders, adds another layer of complexity to its symptomatology (McGrath et al., 2020). Large epidemiological studies show that 40 - 70% of people with depressive disorders also meet criteria for at least one type of anxiety disorder (Lamers et al., 2011; Kessler et al.,

2015), and anxiety symptoms have been identified as a risk factor for poorer treatment outcomes (Fava et al., 2008) depression relapse (Buckman et al., 2018).

A third layer of complexity lies at the individual level of depression, as many depressive symptoms are highly dynamic over time. Studies examining the dynamics of daily affect in major depressive disorder show that not only is negative affect the prevailing affect in depression, but depressed patients also experience greater emotional variability (i.e., larger mood shifts), particularly in negative affect, greater emotional inertia (i.e., slower mood shifts), and greater reactivity of negative affect to daily life events (Nelson et al., 2018). In addition, different patterns of combination between average levels of affect and mood variability have been found in depressed patients (van Genugten et al., 2022). High fluctuations between daytimes have also been found for rumination, with depressed individuals engaging in higher levels of rumination in the morning and evening compared to the middle of the day (Takano & Tanno, 2011). These findings demonstrate that depressed mood and cognitive processes of depression, such as rumination, are highly variable within days and across days. This complexity of depression, with its diverse symptom profiles, frequent comorbidities, and temporal dynamics, poses a significant challenge to both its reliable assessment and effective treatment.

The role of Cognitive Processes in Depression

Although dysfunctional cognitive processes are not explicitly included in the diagnostic criteria of depression (WHO, 2022), they play an important role in the development and maintenance of depression and are therefore an important target for psychological interventions (Garratt et al., 2007). According to cognitive theories of depression, dysfunctional cognitions are the central drivers of depression. For example, Beck's cognitive theory of depression (Beck, 1979) suggests that individuals with depression hold dysfunctional beliefs about themselves, the future and the world that shape their interpretation of stimuli and trigger negative automatic thoughts. These cognitions can perpetuate and exacerbate the symptoms of depression by setting in motion vicious cycles of dysfunctional cognitions, emotional, physiological and behavioral reactions, and their external consequences.

Repetitive Negative Thinking (RNT) – a central Cognitive Process in Depression

In addition to the dynamic between dysfunctional beliefs and negative automatic thoughts, repetitive negative thinking (RNT) has been proposed as a central cognitive process that plays an important role in the onset and maintenance of depression (Ehring & Watkins, 2008). Research on RNT in depression is particularly promising because RNT occurs transdiagnostically in several emotional disorders (Ehring & Watkins, 2008; Wahl et al., 2019; Egan et al., 2024), including anxiety disorders such as general anxiety disorder (GAD), which occurs highly comorbid with depression (McGrath et al., 2020). RNT is the cognitive process shared by rumination and worry that is repetitive, intrusive, relatively uncontrollable and of negative content (Ehring & Watkins, 2008). Conceptually rumination and worry differ in the temporal reference of their content. While rumination is more focused on the past and one's own distress (Nolen-Hoeksema, 1991, e.g., "I should have handled that argument differently, I'm such a failure."), worry is more focused on the future and on potential negative outcomes (Borkovec et al., 1983, e.g., „What if I lose my job and can't find another one? How will I pay my bills?"). People's engagement in RNT is proposed to be a maladaptive mental strategy to cope with emotional distress (Hong, 2007). Worry is proposed to be a form of cognitive avoidance in which people attempt to reduce negative emotions by mentally preparing for anticipated negative outcomes (Borkovec et al., 1998), and rumination is seen as a response style to depressed mood in which people attempt to understand the nature and implications of their negative feelings (Nolen-Hoeksema, 1998). In these ways, RNT inhibits the use of other, more adaptive forms of coping, such as task-oriented problem solving (Hong, 2007).

As RNT is open to intervention (Spinhoven et al., 2018), it has become a specific focus of some psychological interventions (Egan et al., 2024), and even interventions that are not specifically focused on RNT have been shown to be effective in reducing RNT. Meta-analyses (Spinhoven et al., 2018; Bell et al., 2023) show that RNT-focused cognitive behavioral therapy (CBT) and traditional CBT are comparably effective in reducing RNT and depression. Surprisingly however, only for RNT-focused CBT the effects on RNT and depression are strongly correlated. A possible explanation is that several CBT techniques that target depressive symptoms also affect RNT. CBT techniques such as thought stopping techniques, behavioral activation (Saberi et al., 2024), and cognitive restructuring, may break vicious cycles of RNT by improving mood and redirecting one's attention away from negative thoughts. This means that while RNT may be a mechanism of change in RNT-

focused treatments, in traditional CBT change in RNT might rather be an epiphenomenon of general therapeutic improvement (Monteregge et al., 2020). The actual mechanisms of psychotherapy on RNT are not yet fully understood.

The lack of Concreteness in RNT

While the associations between the process features of RNT and depression have been well studied, there is less consensus about the relevance of specific content features of RNT. For instance, studies have shown that the temporal reference of RNT (i.e., whether someone ruminates about the past or worries about the future) is differentially associated with depression and anxiety (Funk et al., 2022), but does not explain additional variance in internalizing symptomatology in general beyond the process of RNT per se (Taylor & Snyder, 2021). However, this does not preclude the possibility that other content features of RNT are uniquely associated with depressive symptoms, and specifically with change in depression. The repetitive nature of RNT implies questions about why RNT thoughts cycle and how these cycles can be disrupted. Moreover, RNT appears to be a process that all people experience from time to time, raising the question of why some people manage to escape from RNT cycles while others remain stuck.

Authors who have extensively researched the content of worry and rumination posit that their maladaptiveness is specifically associated with a reduced concreteness of the content (Joormann et al., 2006; Stöber & Borkovec, 2002). In the reduced-concreteness theory, Stöber & Borkovec (2002) describe that the concreteness of thoughts can range from concrete, i.e., “distinct, situationally specific, unequivocal, clear, singular” (e.g., ‘I didn't study enough to pass the exam.’) to abstract, i.e., “indistinct, cross-situational, equivocal, unclear, aggregated” (e.g., ‘I am a failure.’). The theory is based on two mechanisms explaining the maintenance of worry: First, undetailed and unspecific elaborations impair the generation of problem solutions, and second, they inhibit the production of images (Borkovec et al., 1998), which are essential for a successful emotional processing of problems (Foa & Kozak, 1986). This explains how persistent worry can develop and persist: when a person worries about a potential problem, the worries are maintained and/or worsen when attempts to solve or cope with the problem are unsuccessful.

Stöber‘s theory and his concreteness scale have been investigated in several studies (Stöber et al., 2000; Goldwin & Behar, 2012; McGowan et al., 2017; Stöber & Borkovec, 2002). For instance, studies of both healthy controls (Stöber et al., 2000) and patients with

GAD (Stöber & Borkovec, 2002) have found that descriptions of problems that individuals repeatedly worry about are more abstract than their descriptions of other problems. Furthermore, the worry descriptions of GAD patients are on average less concrete than those of healthy controls, and with successful CBT, the descriptions of GAD patients become more concrete. In parallel, reduced concreteness has also been investigated in rumination and depression (e.g., Watkins & Moulds, 2005a; Watkins & Moulds, 2007). Using the same operationalization as Stöber & Borkovec (2002), Watkins & Moulds (2007) examined reduced concreteness in rumination. Building on Ströber's theory (Stöber & Borkovec, 2002), they investigated that highly abstract and analytical forms of RNT are particularly associated with depression (as opposed to more concrete, experiential forms of RNT). It is important to note that there has been a contradictory finding by other researchers (Kircanski et al., 2015), but this study used a different methodology. When comparing RNT processes in depression, GAD, and healthy controls, the study did not find that clinical groups rated RNT episodes as more abstract than community controls. However, while Stöber & Borkovec (2002) and Watkins & Moulds (2007) used independent ratings of abstractness with the Stöber Concreteness Scale (Stöber & Borkovec, 2002), Kircanski et al. (2015) used self-reports and reported that ratings of abstractness were generally low in their sample. As individuals vary in the degree to which they are aware of the functions of RNT, Wahl et al. (2019) pointed out that self-reports such as those used by Kircanski et al. (2015) reveal more about patients' perceived concreteness of RNT rather than objective ratings. The use of external ratings by trained raters is therefore proposed to be an important aspect to ensure reliable assessments of the concreteness of RNT.

Ecological Momentary Assessment – Studying depression in real-time

Currently, 4.5 billion people, i.e. almost 70% of the world's population, own a smartphone, and it is estimated that the number will peak at 6.2 billion users by 2029 (Statista, 2024). As a result of the increased availability of smartphones and wearable devices, a growing number of studies have begun to explore the potential of mobile applications for clinical research and practice (for a review, see: Colombo, Fernández-Álvarez et al., 2019). Ecological Momentary Assessment (EMA), also known as Experience Sampling Method (ESM), is increasingly being used to study mental health conditions in real time, offering significant opportunities to improve our assessment, treatment and understanding of depression.

Opportunities for a more reliable Assessment of Depression

Traditionally, the clinical assessment of depression relies on patients‘ retrospective self-report of symptoms experienced over the past week or weeks (e.g., Hautzinger et al., 2009; Kroenke et al., 2001). The report is either assessed indirectly, i.e. patients are interviewed by a clinician, or directly through self-report questionnaires (Bondolfi et al., 2010). Clinical interviews are considered the gold standard for assessing depression because they include the clinician's evaluation (Stuart et al., 2014). However, as questionnaires are less resource- and time-consuming they are more practical, especially for monitoring change in depression severity over multiple time points (Bondolfi et al., 2010). In clinical trials, depression questionnaires are typically administered before and after treatment or on a weekly basis, to track changes occurring throughout therapy and to draw conclusions about treatment effects.

As affective symptoms of depression and cognitive processes such as RNT are highly variable within and between days (Wirz-Justice, 2008), the reliability of their retrospective recall is controversial. As Targum (2020) claims, it is „unrealistic to presume that a single point in time measurement can accurately and reliably capture the true symptom severity“ experienced by depressed patients over a past week or weeks. Despite the high fluctuations of certain depressive symptoms and cognitive processes in depression, such as RNT (van Genugten et al., 2022; Rosenkranz et al., 2020), depression is associated with increased cognitive biases that affect the memory and recall of emotional experiences (Gorin & Stone, 2001). Several studies have found that people tend to overestimate their experiences of positive and negative affect when retrospectively recalling them (Ben-Zeev et al., 2009; Colombo, Suso-Ribera et al., 2019; Kardum & Daskijević, 2001; Wirtz et al., 2003). In general, this recall bias occurs in both depressed and healthy individuals, but in depression it is amplified and particularly pronounced regarding the recall of negative affect (Colombo et al., 2020; Gotlib & Joormann, 2010). An EMA study involving 47 healthy individuals (Colombo, Suso-Ribera et al., 2019) compared two weeks of momentary self-reports of positive and negative affect with retrospective self-reports assessed with the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) and examined potential recall biases in interaction with mild depressive symptoms. As hypothesized, participants with higher depressive symptoms showed a greater overestimation of negative affect and a greater underestimation of positive affect, whereas participants with lower or no depressive symptoms showed opposite effects, overestimating their positive affect and underestimating

their negative affect in the retrospective recall. This supports the assumption that retrospective depression questionnaires are biased by recall and suggests that EMA captures the temporal dynamic of depression more reliably.

There is also preliminary empirical evidence for the validity of EMA in predicting clinical interview outcomes. A clinical study showed that changes in depression assessed continuously with EMA over a 6-week antidepressants trial significantly predicted changes in clinician-rated depression before and after the intervention (Targum et al., 2021). However, several aspects of the data quality of EMA require further research. For example, it is important to examine the validity of EMA in predicting global intervention outcomes rated by clinicians, such as global functioning, compared with the predictive validity of established clinical questionnaires.

Opportunities for the Development of Personalized Treatments for Depression

The complexity of depression poses significant challenges for the assessment of depression, but also for its treatment. Even with evidence-based treatments such as psychotherapy and pharmacotherapy, the treatment of depression remains a major challenge. As described above, meta-analyses reveal that only about 40% of patients respond to first-line treatments, and remission rates are even lower (Cuijpers, Karyotaki et al., 2021). In addition, long waiting times for psychotherapy remain a global problem. In many European countries and the United States, patients face an average waiting time of more than 3 months before receiving psychotherapeutic treatment (Barbato et al., 2016; Peipert et al., 2022; Friederich et al., 2024). Long waiting times for treatment not only prolong patient suffering, but are also associated with poorer treatment outcomes (Ghio et al., 2014; van Dijk et al., 2023). Given the high costs associated with psychotherapy and the substantial economic burden of depression in terms of ill days and lost productivity, improving the effectiveness of treatments and their efficient distribution is also a critical priority for healthcare systems (Arias et al., 2022; König et al., 2020). As a result, healthcare systems are increasingly adopting stepped care approaches, in which the level and intensity of treatment is adapted to the severity of patients' symptoms and their response to previous treatments, in order to allocate limited and expensive therapeutic resources more efficiently (van Straten et al., 2015). For example, for mild depressive symptoms, the German National Disease Management Guideline on Unipolar Depression (Bundesärztekammer [BÄK] et al., 2022) recommends close monitoring of symptoms and physician-supervised treatment with low-threshold psychoeducational self-

help or internet- and mobile-based interventions (IMIs), whereas psychotherapy and/or medication are recommended only when symptoms stagnate, worsen, or recur. In contrast, the guidelines recommend immediate psychotherapy or medication for moderate depressive symptoms and a combination of both for severe depressive symptoms. In addition, the guidelines offer recommendations on how to proceed with non-responders. For instance, if a patient does not respond to psychotherapy, it is recommended to evaluate the patient-therapist relationship and the applied therapeutic approach. Depending on the evaluation, therapy may than be intensified (e.g., increasing the frequency of sessions) or the patient may be referred to another provider.

The low response rates of depression treatments are often attributed to the heterogeneity of depression and the variability of individual responses (e.g., Fried, 2017; Simmonds-Buckley et al., 2021; Cohen & DeRubeis, 2018). That is, although interventions have moderate effects on average, treatment responses are highly variable at the individual level (Cuijpers, Oud et al., 2021). A large part of treatment research therefore focuses on the development of "personalized", sometimes also referred to as "precision" medicine, in which the choice of treatment (e.g., DeRubeis et al., 2014; Friedl et al., 2020) or the combination of treatment modules (e.g., Fisher et al., 2019) is tailored to patients'individual needs (for a review, see: Cuijpers et al., 2016).

The development of personalized treatment approaches poses major challenges (for a review, see: Lorenzo-Luaces et al., 2021). Designing personalized treatments requires assigning different therapeutic approaches or combining different therapeutic modules in ways that specifically fit patients' individual profiles (DeRubeis et al., 2014; Herpertz & Schramm, 2022). However, at the individual level (i.e., within-person), it is not feasible to test the superiority of one treatment over another because the same person cannot be treated with two different treatments without introducing significant confounding factors, such as the effect of the treatment order (Lorenzo-Luaces et al., 2021). Therefore, individual profiles need to be grouped into subgroups so that the effectiveness of different treatments can be tested between individuals. In personalized modularized treatments, this means that even without considering the sequence and dosage of individual modules, all possible combinations of modules must be tested for their effectiveness to determine the most effective combination for a specific subgroup (Herpertz & Schramm, 2022). However, there are endless possibilities for building subgroups. As already lined out, the diagnosis of depression encompasses several thousand possible symptom profiles (Fried & Nesse, 2015), and besides symptoms, other

factors (e.g., sociodemographic variables auch as age or gender) are considered as predictors of treatment response (Kessler et al., 2017). Therefore, to achieve reliable and generalizable results, researchers need to conduct large-scale randomized controlled trials (RCTs) with sufficient statistical power to detect (most likely small) differences between most effective and second most effective treatment approaches (Lorenzo-Luaces et al., 2021). As such studies are very time-consuming and require substantial resources (Blackwell et al., 2019), strategies that can speed up these developments are highly demanded (Blackwell et al., 2019; Huibers et al., 2021; Lorenzo-Luaces et al., 2021). Two strategies that could accelerate the development of personalized treatments are: the improvement of outcome reliability and the early prediction of the treatment outcome.

In depression, the intervention effects of clinical trials are typically assessed with retrospective questionnaires. As outlined above, the use of EMA could increase the reliability of clinical assessments. A simulation study (Schuster et al., 2020) shows that such an increase would positively affect the power of statistical analyses and could therefore reduce the sample sizes needed to detect significant intervention effects between intervention conditions. A first empirical study (Moore et al., 2016) supports this assumption. They investigated the effects of mindfulness therapy on mindfulness, depression, and anxiety with EMA and retrospective questionnaires. In line with the hypothesis, they found higher intervention effects for EMA-assessed mindfulness and depression compared to questionnaire-assessed measures. Accordingly, the resulting number-needed-to-treat was 25-50% lower for the EMA than the questionnaire outcome. This means in their modest sample size of around thirty patients per condition, Moore et al. (2016) found significant intervention effects with EMA that were not (yet?) detected by the questionnaire assessments. In clinical trials, this could enable a more rapid identification of effective treatments and/or reduce the sample sizes required to achieve enough power for the detection of intervention effects.

To establish EMA as an alternative assessment technique to questionnaires in clinical trials, these initial findings need to be externally validated. For instance, it is unclear whether EMA is still more sensitive to intervention effects when the compared questionnaire assessments are administered weekly rather than just before and after the intervention. Moreover, it is important to examine whether the results are stable in an EMA approach that is shorter than the questionnaire, which is typically required to reduce the patient burden induced by frequent assessments.

A second strategy that could speed up the evaluation of the effectiveness of interventions is to investigate factors that predict intervention outcomes early thereon. In pharmacological treatments this strategy is already used. Systematic research on early treatment prediction in pharmacotherapy for depression (Szegedi et al., 2009) has led to clear guidelines for the adjustment of medication. Specifically, it is recommended to consider medication or dosage change if no improvement is achieved after four weeks (Gautam et al., 2017; Bundesärztekammer [BÄK] et al., 2022).

For psychological interventions, early treatment prediction could not only speed up clinical trials, but also lead to better clinical outcomes by reducing the time patients‘ spend in ineffective treatments before considering alternatives (Schaffer et al., 2013), and moreover improve the efficient distribution of scarce therapeutic resources in terms of stepped care (e.g., considering an intensification of the treatment; Richards, 2012). However, such guidelines are lacking for psychological interventions. Indeed, there is compelling evidence for the predictive value of early improvements in the outcomes of psychological interventions, but the existing studies are highly heterogeneous and, in particular, lack consensus on which time window and rate of improvement early in the treatment is the most predictive for distinguishing between treatment responders and non-responders (Beard & Delgadillo, 2019; Li et al., 2023). This means that existing studies have examined different time windows classified as “early” (e.g., two, four, six or eight weeks after baseline) and different change criteria classified as an “improvement” (e.g., change rates, reliable or clinically significant improvement or the occurrence of sudden gains; Gois et al., 2014; Rubel et al., 2015; Hunnicutt-Ferguson et al., 2012). Furthermore, only one study, specifically an online intervention study (Schibbye et al., 2014), has investigated the potential of EMA for assessing early predictive change.

As EMA integrates into patients‘ daily lives and can frequetly assess early change, it is a promising technique for frequent treatment monitoring. Therefore, it is important to further explore the predictive validity of EMA for early treatment response prediction in psychological treatments and to investigate which time window and change criteria are most predictive of the treatment outcome.

Opportunities for a better Understanding of Depression

Another opportunity for EMA in psychotherapy research is its potential to improve our understanding of psychopathology. As smartphones have become people‘s daily companions

(Statista, 2023), EMA can provide ecological insights into people‘s daily feelings, thoughts and behaviors, their temporal dynamics (e.g., Kircanski et al., 2018), and how they change over the course of treatment (e.g., Goodman et al., 2023). For instance, it allows for the investigation of the temporal dynamics between depressive symptoms and cognitive processes such as RNT, as well as whether and how their dynamics change over the course of psychotherapy.

While in the past the dynamics between rumination, worry, and depression have been studied mainly in the laboratory, showing, for example, that experimentally induced rumination increases depressed mood (Nolen-Hoeksema & Morrow, 1993), impairs problem solving (Lyubomirsky & Nolen-Hoeksema, 1995; Watkins & Moulds, 2005a), and interferes with instrumental behavior (Nolen-Hoeksema et al., 2008), EMA allows for repeated sampling of RNT at the moment of its natural occurrence (Colombo, Fernández-Álvarez et al., 2019), thereby increasing the external validity of results. As EMA data typically have a hierarchical structure (with occasions being nested within individuals), it allows the use of time series analyses, such as multilevel modelling (MLM), which improves the reliability of results by accounting for within-person and between-person variability (Snijders & Bosker, 2011). Futhermore, by using cross-lagged models, EMA facilitates the investigation of temporal relationships between variables (Hamaker et al., 2015), such as between RNT and depressive symptoms.

For example, EMA studies have found that a) depression is associated with more frequent experiences of stressful events (Moberly & Watkins, 2008; Ruscio et al., 2015), b) experiences of stressful events are followed by increased momentary levels of negative affect (Moberly & Watkins, 2008), rumination (Ruscio et al., 2015; Kircanski et al., 2018), but not worry (Kircanski et al., 2018), c) increases in negative affect following stressful events are partially mediated by rumination (Ruscio et al., 2015; Moberly & Watkins, 2008), d) trait rumination, but not depression, is associated with increased mood reactivity to stressful events (Moberly & Watkins, 2008), e) higher levels of rumination, but not worry (Kircanski et al., 2018) at a given time point predict increases in negative affect and decreases in positive affect (Ruscio et al., 2015; Kircanski et al., 2018), as well as increases in momentary depression and GAD symptoms at the subsequent time point (Ruscio et al., 2015), f) the effect of momentary rumination on subsequent shifts in negative affect, as well as on momentary depression and GAD symptoms, is higher in patients with depression and/or GAD (Ruscio et al., 2015) than in healthy individuals, and g) the effect of momentary rumination on subsequent shifts in

negative affect also exists vice versa, which means higher levels in negative affect also predict greater increases of rumination at a subsequent time point (Moberly & Watkins, 2008; Blanke et al., 2022). The EMA frequency in these studies was six to ten times per day.

Despite the substantial number of studies that investigated momentary levels of worry and rumination in depressed and healthy individuals, only one study investigated the concreteness of RNT on a momentary level (Kircanski et al., 2015). However, this study examined self-reports of RNT concreteness rather than ratings from trained external raters, which as outlined above reveals more about patients‘ perceived concreteness of RNT rather than objective ratings (Wahl et al., 2019). Therefore, it is important to examine whether the concreteness of momentary RNT is a particular mode of RNT that uniquely contributes to the prediction of momentary levels of depressive symptoms beyond the process of RNT per se. If so, it would could be important for future studies to assess the concreteness of RNT in addition to the distinct content-independent process features of RNT (i.e., repetitiveness, intrusiveness, and difficulty to control; Ehring et al., 2011) to increase our understanding of RNT and depression. In addition, the concreteness of momentary RNT and its dynamics with the experience of depressive symptoms may change over the course of psychological treatment. Finally, research on the temporal relationship between RNT concreteness and momentary depressive symptoms, which has not yet been conducted, could provide valuable insights into whether the just-in-time treatment of momentary RNT concreteness might be a promising strategy for reducing momentary depressive symptoms and/or vice versa.

Objectives of the present thesis

This dissertation investigates the opportunities of EMA to support the outlined challenges of psychotherapy research in depression. For the investigation, we conducted an EMA study as part of a larger RCT evaluating the effectiveness of three psychotherapy approaches for depression: cognitive behavioral therapy (CBT), schema therapy (ST), and individual supportive therapy (IST; Kopf-Beck et al., 2024). The trials was conducted at the Max Planck Institute of Psychiatry in Munich, Germany, and included 106 moderately to severely depressed patients, representing about one-third of the total RCT sample (N = 300). Participants were randomly assigned to one of the three interventions, which consisted of two group sessions (100 minutes each) and two individual sessions (50 minutes each) per week for seven weeks. The treatments were delivered in either an inpatient or day clinic setting, alongside standard psychiatric treatments such as pharmacotherapy and complementary

therapies like ergotherapy and case management (Kopf-Beck et al., 2020). In addition to a comprehensive assessment battery that included weekly questionnaire assessments (WQA) of depressive symptoms and RNT, as well as pre- and post-intervention clinical interviews of global functioning, EMA collected momentary self-reports three times daily throughout the entire intervention period, covering the following variables: 'momentary depressive symptoms' (a sum score of four Likert-scaled items including depressed mood, loss of interest, withdrawal, and psychomotor agitation or inhibition), 'momentary RNT' (a sum score of four Likert-scaled items measuring the repetitiveness, intrusiveness and difficulty to control RNT, as well as the perceived burden through RNT), and RNT thoughts assessed via a free-text item. The wordings of the EMA items are presented in Table 1.1.

Table 1.1

Original wordings and English Translation of the EMA items

Variable/ Item	**Original german item wording**	**Englisch translation**
Momentary depressive symptoms – sum score of four items		
Loss of interest[a]	Hast du gerade das Gefühl, zu nichts mehr Lust zu haben?	Do you feel like you don't want to do anything anymore?
Withdrawal[a]	Ziehst du dich gerade von sozialen Kontakten oder Aktivitäten zurück?	Are you currently withdrawing from social contacts or activities?
Psychomotor agitation /inhibition[a]	Fühlst du dich gerade besonders körperlich gehemmt oder aktiviert?	Are you feeling particularly physically inhibited or agitated?
Current mood[b]	Wie fühlst du dich?	How are you feeling?
Momentary Repetitive Negative Thinking (RNT) – sum scores of four items		
Repetitiveness of RNT[a]	Dieselben negative Gedanken gehen mir immer und immer wieder durch den Kopf.	The same negative thoughts keep going through my mind again and again.
Uncontrollability of RNT[a]	Ich hänge an bestimmten negativen Gedanken fest und kann mich nicht davon lösen.	I get stuck on certain negative issues and can’t move on.
Intrusiveness RNT[a]	Negative Gedanken tauchen auf, ohne dass ich dies will.	Negative thoughts come to my mind without me wanting them to.
Subjective burden through RNT[a]	Ich fühle mich durch negative Gedanken beeinträchtigt.	I feel weighted down by negative thoughts.
RNT thoughts – free-text item (the concreteness of the thoughts was later rated by trained external raters)	Welche negativen Gedanken gehen dir aktuell wiederholt durch den Kopf? Bitte schreibe deine Gedanken in ganzen Sätzen auf.	Which negative thoughts are currently going through your mind repeatedly? Please write down your thoughts in complete sentences.

Note. EMA: Ecological Momentary Assessment; [a]The response scale of the EMA Items, except for the mood and the free-text item, was two-stepped. Participants responded to a binary *Ja-Nein* (engl.: *Yes-No*) scale. If *Ja* (engl.: *Yes*) was selected, a five-point Likert scale followed, which assessed the extent of agreement (labeling: *Gar nicht, Ein bisschen, Einigermaßen, Erheblich, Äußerst;* engl.: *not at all, a bit, moderately, considerably, very much*). [b]The 'Current mood' item was rated by selecting one of five emojis (labeling: *Sehr gut, Gut, Mittelmäßig, Schlecht, Sehr schlecht*; engl.: *very good, good, moderate, bad, very bad*).

Based on the collected data, three distinct studies and research articles were worked out, each exploring the potential of EMA to enhance our assessment, treatment, or understanding of depression.

Study I

Despite the rapid rise of EMA in depression research, only a small number of studies have empirically examined how EMA results compare to those from traditional clinical assessment techniques, such as questionnaires and clinical interviews in clinical trials (Moore et al., 2016; Targum et al., 2021). An empirical investigation of their comparability builds an important basis for future research. Methodological differences between EMA, questionnaires and clinical interviews are ususally the number and selection of items, as well as the temporal reference of the assessments (momentary vs. retrospective). Previous studies suggest that EMA might assess symptom change in depression more reliable and be more sensitive to the detection of intervention effects in clinical trials. However, only one study (Moore et al., 2016) has empirically tested this, and there remains a lack of comparative research on the validity of EMA and questionnaires in predicting clinical interview outcomes (Targum et al., 2021).

Therefore, in Study I, we compared the results of EMA versus WQA on change in depressive symptoms and RNT and investigated: (1) the size of the intervention effects associated with both techniques and (2) their validity in predicting clinical interview outcomes of change in global functioning.

Study II

In the pursuit of developing personalized treatments, many researchers have explored factors predicting individual responses to specific psychological interventions. However, these identified factors often lack external validity, meaning they are rarely applicable across

different populations (Lorenzo-Luaces et al., 2021). Nonetheless, one factor has shown consistent predictive power across multiple studies and populations: early improvements in therapy (Beard & Delgadillo, 2019; Li et al., 2023). In pharmacotherapy the predictive value of early improvements has led to clear guidelines for the adjustment of medications (Gautam et al., 2017). However, for psychological interventions such guidelines are lacking, even though they could promote better clinical outcomes (Schaffer et al., 2013) by supporting the adjustments of interventions according to stepped care (van Straten et al., 2015), distribute limited therapeutic resources more efficiently (Richards, 2012), and potentially speed up the evaluation process of clinical trials on personalized therapies (Kidwell & Almirall, 2023). As EMA integrates into patients' everyday lives and is capable of frequent assessments of early change, it might be a promising technique for early treatment prediction research. However, despite strong evidence supporting early improvement as a robust predictor of treatment response in psychological interventions (Beard & Delgadillo, 2019; Li et al., 2023), there remains a lack of consensus on the optimal timing and change rate that serves as the best predictor for early treatment response prediction (Beard & Delgadillo, 2019).

Therefore, in Study II, we investigated three distinct research questions: (1a) At which time point after treatment initiation does early improvement in depressive symptoms significantly predict treatment response? (1b) Do both, WQA and EMA of early improvement significantly predict treatment response at these time points? (2) How predictive are different definitions of early improvement in terms of the defined time window and symptom cutoff, as assessed by EMA versus WQA (we investigated the definitions: ≥10%, 20%, 30%, or 40% improvement after one, two, three, or four weeks of treatment)?

Study III

Finally, we investigated the potential of EMA for the ecological research on cognitive processes, specifically RNT in depression. Unlike previous EMA studies on RNT, we focused specifically on the concreteness (Stöber & Borkovec, 2002) of RNT by having patients journal their RNT thoughts three times daily, which were then rated for concreteness by trained external raters. As introduced, the subprocesses of RNT, rumination and worry are assumed to be not inherently maladaptive, as they may be necessary for problem-solving (Joormann et al., 2006; Stöber & Borkovec, 2002). However, studies show that rumination and worry thoughts of depressed patients are less concrete compared to healthy individuals, which means they are more unclear, aggregated, cross-situational and less solution-oriented (Stöber & Borkovec, 2002). While it is well studies that psychotherapeutic approaches, such

as CBT, effectively reduce RNT (Bell et al., 2023), changes in RNT concreteness during psychotheapy have not yet been investigated on a momentary level. Moreover, the temporal dynamics of momentary depressive symptoms have only been studied with momentary levels of RNT, but not with the concreteness of momentary RNT, which is proposed to moderate the effect of RNT on depressed mood (Watkins & Moulds, 2005a).

Therefore, the third study investigated three distinct research questions: (1) Does the concreteness as a particular mode of momentary RNT explain variance in the prediction of momentary depressive symptoms beyond the process of momentary RNT per se? (2) Does the concreteness of momentary RNT increase over the course of psychotherapy? (3) How are momentary depressive symptoms associated with the concreteness of momentary RNT, does one factor temporally precede the other?

2. Study I:

Ecological Momentary Assessment versus Weekly Questionnaire Assessment of Change in Depression

This chapter is a post-peer-review, pre-copyedit version of an article published in *Depression and Anxiety*.

Tamm, J., Takano, K., Just, L., Ehring, T., Rosenkranz, T. & Kopf-Beck, J. (2024) Ecological Momentary Assessment versus Weekly Questionnaire Assessment of Change in Depression. *Depression and Anxiety*, 13 (1), 9191823.

The final authenticated version is available online at:
https://doi.org/10.1155/2024/9191823

Abstract

Objective: Ecological momentary assessment (EMA) is increasingly used to monitor depressive symptoms in clinical trials, but little is known about the comparability of its outcomes to those of clinical interviews and questionnaires. In our study, we administered EMA and questionnaires to measure change in depressive symptoms and repetitive negative thinking (RNT) in a clinical trial and investigated a) the size of intervention effects associated with both techniques and b) their validity in predicting clinical interview outcomes (i.e., global functioning). **Method**: Seventy-one depressed patients were randomly assigned to one of three psychological interventions. The EMA comprised a concise item set (4 items per scale) and was administered three times per day during a seven-week intervention period. Conversely, questionnaires were assessed weekly (WQA), encompassing their full sets of items of depressive symptoms and RNT. **Results**: While EMA excelled in detecting significant intervention effects, WQA demonstrated greater strength in predicting clinician ratings of global functioning. Additionally, we observed significant differences in time effects (slopes) between the two techniques. WQA scores decreased steeper over time and were more extreme, e.g., higher at baseline and lower post-intervention, than EMA scores. **Conclusions**: Although clinical interviews, questionnaires and EMA outcomes are related, they assess changes in depression differently. EMA may be more sensitive to intervention effects, but all three methods harbor potential bias, raising validity and reliability questions. Therefore, to enhance the validity and reliability of clinical trial assessments, we emphasize the importance of EMA approaches that combine subjective self-reports with objectively measured behavioral markers.

Introduction

Accurate reporting of changes in depression is essential for symptom monitoring and research on the effectiveness of interventions. Retrospective questionnaires, commonly used for this purpose, are usually administered pre- and post-intervention or on a weekly basis. Consequently, they average symptom severity based on patients' recall over the past week or weeks (Hautzinger et al., 2009; Kroenke et al., 2001). Despite their common use, the validity of retrospective questionnaires is questionable. Depression is a dynamic disorder with large symptom fluctuations over time (Wirz-Justice, 2008). Particularly, depressed mood and processes of depression such as repetitive negative thinking (RNT) are known to be highly variable within days and across multiple days (Chen et al., 2022; Peeters et al., 2006; Takano & Tanno, 2011). RNT describes the cognitive process of recurrent dwelling on negative content, often experienced as intrusive and challenging to control. It includes worry and rumination and is regarded as a transdiagnostic process that plays a central role in the development and maintenance of emotional disorders (Ehring & Watkins, 2008). Given the pronounced fluctuations in symptoms and associated processes like mood and RNT, retrospective measurement is challenging. The reason lies in the human memory of emotional experiences, which is distorted, especially in depression (Gorin & Stone, 2001). Numerous studies have investigated that people tend to overestimate their experiences of positive and negative affect when asked to recall them retrospectively (Ben-Zeev et al., 2009; Colombo, Suso-Ribera et al., 2019; Kardum & Daskijević, 2001; Wirtz et al., 2003), and in depression this recall bias is further amplified, particularly concerning negative affect (Colombo et al., 2020; Gotlib & Joormann, 2010).

An alternative assessment technique that prevents recall bias is ecological momentary assessment (EMA; Ebner-Priemer & Trull, 2009a; Moskowitz & Young, 2006; Trull & Ebner-Priemer, 2009). Conducted on smartphones, EMA takes place in patients' daily lives and allows repeated sampling of psychological states such as feelings, thoughts, or behavior in the moment (Shiffman et al., 2008). Although EMA and other time-series-based procedures are increasingly used in clinical research, the practice of implementing them into clinical trials is not widespread (Colombo, Fernández-Álvarez et al., 2019). Moore et al. (2016) and Targum et al. (2021) were the first clinical trials tracking depressive symptoms with EMA. Moore et al. (2016) examined the intervention effects of mindfulness therapy on depression, mindfulness and anxiety using EMA administered 10 days before and after the intervention compared to point-assessments with questionnaires. They investigated that EMA was associated with more pronounced intervention effects for mindfulness and depression.

Moreover, the Number Needed to Treat (NNT) was 45% to 74% lower for the EMA compared to the questionnaire assessments. The authors conclude that EMA may be more sensitive in detecting and quantifying intervention effects due to its avoidance of recall bias inherent in retrospective questionnaires. Additionally, Targum et al. (2021) discovered that change in depression assessed with EMA continuously over a 6-week antidepressants trial predicted change in depression rated by clinicians pre- and post-intervention.

Previous studies showed that EMA is an efficient and valid assessment technique to assess change in depression and they support the hypothesis that EMA is more sensitive to this change than point-assessments with questionnaires. However, to establish EMA as an alternative assessment technique to questionnaires in clinical trials, further investigations are needed. Questions, that need more research are: How comparable are EMA and questionnaires assessments a) when questionnaires are administered weekly instead of just before and after the intervention, b) when the EMA is more brief than the questionnaire, which is typically a need in order to reduce participants' burden when responding to several occasions per day and c) when comparing their validity in predicting global intervention outcomes rated by clinicians, such as global functioning?

In this study, we measured change in depressive symptoms and RNT with EMA continuously over the course of a seven-week clinical trial in comparison to weekly questionnaire assessments (WQA). The aim was to investigate two different aspects. First, our aim was to replicate the findings of Moore et al. (2016) by testing whether EMA demonstrates larger intervention effects in the comparison of two different intervention conditions than questionnaires. In line with Moore et al. (2016), we hypothesized that change in depressive symptoms and RNT assessed with EMA would be associated with larger intervention effects. Second, we investigated differences between the two assessment techniques in predicting change in global functioning rated by clinicians. We hypothesized that, after controlling for baseline global functioning, changes in depressive symptoms and RNT assessed with EMA would predict post-intervention global functioning more strongly than the same predictors assessed with WQA. For our analysis, we used data from the OPTIMA study (Kopf-Beck et al., 2020), a clinical trial investigating the effectiveness of three different psychological interventions for depression: schema therapy (ST) versus individual supportive therapy (IST) and cognitive behavioral therapy (CBT) administered to moderately to severely depressed patients in an inpatient and day clinic setting.

Materials and Methods

Participants

To design our sample size, we performed computer simulations to detect a significant interaction between time and the intervention conditions with multilevel models (MLM). We based these simulations on the findings of Moore et al. (2016), where the mindfulness-based stress reduction intervention had an effect of Cohen's d = 0.4 on the mindfulness outcome at the post-intervention assessment. The results of our simulations indicated that the required sample size is around 20 per condition (resulting in 60 patients in total) to achieve a power of 0.80 under alpha = 0.05. We calculated with a mean drop-out rate of 20% (experiences of the OPTIMA study) and additional 20% due to insufficient EMA data, so that we aimed a sample size of n = 33 patients per intervention condition.

Drop-outs during the conduct of the study were defined as enrolled patients who were found to have incorrect in- or exclusion criteria during the conduct of the study, who left the clinic before end of intervention, or who missed more than six sessions (22%) of their intervention. Survival analyses were conducted to test for differences in drop-out risk between our intervention conditions.

Design and Procedures

We analyzed data collected as part of the OPTIMA study (Kopf-Beck et al., 2020; Identifier on clinicaltrials.gov: NCT03287362), a monocentric, rater-blinded, prospective, parallel-group, block-randomized clinical trial with repeated measures and three intervention conditions (CBT, ST and IST). The OPTIMA study was conducted at the Max Planck Institute of Psychiatry in Munich, Germany.

Inclusion criteria were age 18 to 65 years, having a compatible smartphone and a diagnosis of a major depressive disorder, single episode or recurrent, moderate or severe without psychotic symptoms diagnosed by clinical assessment. Patients received an expense allowance based on the response rate they achieved in the EMA. The study protocol was approved by the Institutional Ethic Committee of the Faculty of Medicine at LMU Munich (Project number 17–395). All participants provided written informed consent prior to clinical interviews, further measures and randomization. More detailed information about exclusion criteria and further procedures of the OPTIMA trial are given elsewhere (Kopf-Beck et al., 2020).

Interventions

Patients enrolled in the OPTIMA study were randomly allocated to one of three intervention conditions (schema therapy (ST), individual supportive therapy (IST) or cognitive behavioral therapy (CBT)), each one lasting seven weeks and consisting of two individual (50 min each) and two group (100 min each) sessions per week. Details about the different interventions are described in the OPTIMA study (Kopf-Beck et al., 2020). CBT, which is based on Beck's theory of depression (Beck, 1979, 2002) is recommended as first-line psychological intervention for depression (Bundesärztekammer [BÄK] et al., 2022). In contrast, ST is a transdiagnostic psychological intervention that is mainly rooted in cognitive therapy but integrates techniques of different therapeutic approaches such as psychodynamic therapy, gestalt therapy, and ergotherapy (Young et al., 2006). IST was used as an active and nonspecific approach that follows the concept of a bio-psycho-social disease model of depression and is based on the common factors of psychotherapy (Frank, 1971; Grawe, 1995; Greenberg, 2004).

Concomitant care

The OPTIMA study design did not regulate parallel psychopharmacotherapy or potential influencing factors inherent to an inpatient or day clinic intervention program, such as ergotherapy or case management. Decisions hereon were left to the psychiatrist in charge. To mitigate biases, all concomitant care was documented for subsequent use as potential confounders in the statistical analysis.

Measures

Comprehensive information regarding all measures conducted in the OPTIMA trial, encompassing various questionnaires, clinical interviews, imaging and tests, is available in the OPTIMA study (Kopf-Beck et al., 2020). Only the measures of the EMA sub-study relevant to our analysis are described here. These include five primary outcome variables: Questionnaire and EMA scores of depressive symptoms and RNT and global functioning measured with a clinical interview pre- and post-intervention. Additionally, a short feasibility questionnaire of the EMA was assessed at the end of the intervention. Table A.1 in the Supplementary Material provides an overview of the assessment plan.

EMA

EMA was conducted continuously throughout the entire intervention period, starting directly after patients' enrolment in the study. It comprised three prompts per day and was signal-contingent, i.e., patients were automatically prompted by their device. Each day was divided into three phases (morning, noon, and evening). When installing the EMA app, participants reported their approximate wake-up time. The times of the three phases were based on this approximate wake up time (morning: two hours before - 5 hours after approximate wake-up time; noon: 5 - 10 hours after the approximate wake-up time; evening: 10 hours after - two hours before the approximate wake-up time). During each phase, patients received one EMA prompt, which could only be completed within its assigned phase. Once a phase was over, the prompt could no longer be completed, but the subsequent phase's prompt was provided for participants to respond. Additionally, patients received semi-randomized reminder to complete the EMA prompts. The mean time between patient responses was M = 303.8 min (SD = 94.06 min) (morning to noon), M = 337.8 min (SD = 106.66 min) (noon to evening) and M = 810.6 min (SD = 110.03 min) (evening to morning). This suggests that the randomization procedure successfully prevented temporal clustering of responses. To control for sequence effects, the EMA item order was randomized across prompts. Including baseline, this resulted in a total of 168 (56 days * 3 prompt per day) EMA prompts.

The EMA score of depressive symptoms was calculated by summing four EMA items, which were developed by the first and the last author in consultation with clinicians and based on diagnostic criteria for major depression (ICD-10). The four items represent three core symptoms of depression (loss of interest, withdrawal and psychomotor agitation/inhibition) and current mood.

The EMA score of RNT was recently developed by another study (Rosenkranz et al., 2020) and has shown excellent model fit, high reliability and good validity with depression outcomes. The paradigm comprises four EMA items. Three of the items are from the Perseverative Thinking Questionnaire (PTQ) representing the core characteristics of RNT (repetitiveness, intrusiveness and uncontrollability). The fourth item measures subjective burden through RNT. Specific wordings of all EMA items can be found in Table A.2 in the Supplementary Material.

We computed internal reliabilities for both EMA total scores (Depression and RNT) using the 'multilevel.reliability()' function in the 'psych' package of R (Cranford et al., 2006;

Revelle, 2022; Revelle & Condon, 2019). This function calculates both within-participant reliability of change over time points (i.e., Rc) and between-participant reliability of the averaged scores over k number of timepoints (i.e., RkF). For both, depressive symptoms and RNT, we observed good within participant reliability (depressive symptoms: Rc =0.79, RNT: Rc = 0.86) and excellent between participant reliability (depressive symptoms: RkF = 0.1, RNT: RkF = 0.1).

The response scale of all EMA Items, except for the mood item, was two-stepped: Participants responded to a binary Yes-No scale (i.e., *The same negative thoughts keep going through my mind again and again, Yes* or *No*). If *Yes* was selected, a five-point Likert scale followed, which assessed the extent of agreement (labeling: *not at all, a bit, moderately, considerably, very much*). In contrast, if *No* was selected, the Likert scale did not appear, and the next item followed. Participants rated their mood by selecting one of five emojis (labeling: *very good*, *good*, *moderate*, *bad*, *very bad*) that best described their current mood.

Weekly Questionnaire Assessments

The corresponding questionnaire scores of depressive symptoms and RNT were assessed weekly, resulting in a total of 8 assessment points, including baseline. Depressive symptoms were assessed with the Beck Depression Inventory (BDI-II; Hautzinger et al., 2009), a widely used self-rating instrument that accounts for different depressive symptoms. RNT was assessed with the Perseverative Thinking Questionnaire (PTQ; Ehring et al., 2011), which evaluated three core characteristics of RNT: repetitiveness, intrusiveness, and the difficulty in disengaging from negative thoughts.

Clinical interviews

Clinical interviews were conducted to assess patients' global functioning. Trained and blinded raters employed the World Health Organization Disability Assessment Schedule (WHO-DAS; Kirchberger et al., 2014). In the OPTIMA study interrater reliability (intraclass coefficient) was routinely assessed, showing excellent agreement (M = 0.998, SD = 0.004; Kopf-Beck et al., 2024).

Statistical Analyses

As outlined in our preregistered analysis, we planned to filter patients with an EMA response rate below 33%. Our intention was to align with similar approaches found in the literature (Moore et al., 2016). However, new methodological recommendations (Jacobson,

2020) suggest setting such arbitrary cutoffs lowers statistical power. Therefore, we decided to deviate from the preregistered protocol, i.e., to include all patients in the analyses regardless of their response rate. As a sensitivity analysis, we repeated the same analyses with the planned cutoff (i.e., to omit data from patients with a response rate < 33%), and we found that the results were unchanged. As planned, however, we checked person-level standard deviations for each EMA-item across the trial period. For plausibility, patients with a standard deviation of zero in at least one EMA item were excluded. Additionally, we recognized some prompts with missing items, which we filtered before running the formal analyses.

Differences between intervention conditions in demographics, baseline variables, and response rates were examined using chi-squared tests for non-parametric variables and ANOVA tests for parametric variables. Baseline depression, gender, response rates in EMA and WQA, intervention condition, and concomitant care were included as covariates in our statistical models. Our primary variables were standardized using individual person's means and standard deviations, allowing for a conversion of EMA and questionnaire values into a common unit. In all our statistical analyses, a p-value less than 0.05 was considered significant.

Hypothesis 1

To test our first hypothesis, we estimated four parallel MLMs, separately for depressive symptoms or RNT assessed with EMA or WQA. The MLMs were implemented using the R package 'nlme' (Pinheiro et al., 2021). MLMs accommodate the nested structure of the data, with occasions (level 1) nested within individuals (level 2). An alternative three-level data structure with occasions nested within days nested within individuals was considered but rejected to keep model complexity low. Adding the day level into a three-level model explained less than 2% of additional variance (Intraclass Correlation (ICC) of day in depressive symptoms: 1.6%, ICC of day in RNT: 1.97%). The two-level MLMs were specified as follows:

$$Y_{ij} = int_j + slp_j * time_{ij} + r_{ij},$$

$$int_j = \gamma_{00} + \gamma_{01}(intervention\ condition_j) + u_{0j},$$

$$slp_j = \gamma_{10} + \gamma_{11}(intervention\ condition_j) + u_{1j}.$$

where Y_{ij} is the EMA or WQA-assessed levels of depressive symptoms or RNT of the *j*-th participant at time *i*. The residual is denoted by $r_{ij.}$ Note that the unit of time was different

between the EMA and WQA models (moment vs. week). Both intercepts (int_j) and slopes (slp_j) were allowed to vary across individuals (random effects). Individual differences in the intercepts and slopes were explained by the intervention conditions, which were explicitly assumed as fixed effects (γ_{01} and γ_{11}). The effect of the intervention conditions on the slope (i.e., $\gamma_{11,}$ time-condition interaction) was of our particular interest, which represents how the change rate in an outcome differed across the three intervention conditions.

To obtain standardized intervention effects, we calculated Cohen's d (Cohen, 1988) and the NNT (Furukawa & Leucht, 2011). Cohen's d effect sizes were calculated for individual slopes (slp_j) using the 'cohensD()' function from the R package 'lsr' (Navarro, 2015). The NNT index represents the number of individuals who need to undergo treatment for one person to benefit from the intervention compared to an alternative intervention or a control condition. A lower NNT is considered favorable as it implies a higher likelihood of benefitting from the intervention. We calculated the NNT using the function 'NNT()' from the R package 'dmetar' (Harrer et al., 2019), assuming that 44% (reference group: IST) or 46% (reference group: CBT) of the reference condition would respond to the intervention. We defined intervention response as a 50% decrease in BDI-II score from pre- to post-intervention, which is a common definition used in depression literature (Rush et al., 2006).

Hypothesis 2

The second hypothesis was tested by submitting the individual change rates (i.e., slopes of time) from the MLMs of Hypothesis 1 into multiple regression models that predicted patients' global functioning post-intervention. Again, we ran four parallel models, one for each slope derived from the MLMs of EMA- or WQA-assessed depressive symptoms or RNT, serving as the focal predictors of the models.

Additionally, we estimated two combined models including both assessment techniques simultaneously. In all models, we controlled for the baseline levels of global functioning. The separate models were conducted to ascertain if both EMA and WQA slopes significantly predicted post-intervention global functioning independently, and the combined models determined the relative strength of both predictors.

In the last step, we conducted model comparisons to assess variations in the explained variance between the separate and combined prediction models. Specifically, we employed two widely recognized information criteria, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC; Anderson & Burnham, 2004).

Additional Analysis

To explore the distinctions between EMA and WQA more deeply, we performed an additional analysis to our pre-registered analysis plan, investigating whether time effects (slopes) significantly differed between the two assessment techniques. To achieve this, we aggregated the EMA scores of depressive symptoms and RNT to weekly means and conducted MLMs, including 'assessment technique' (EMA vs. WQA) as a predictor. We conducted two separate models, one with depressive symptoms and one with RNT as the dependent variable. The models included occasions and assessment techniques as fixed effects, while considering patients as random effects to account for individual variability. Our primary focus was the assessment technique-by-time interaction as the fixed effect of interest. Given that our EMA and WQA scores predominantly comprised different items, we also conducted corresponding MLMs at the level of three individual RNT items that were identical between the EMA and WQA scores of RNT.

Transparency and Oppeness

We report how we determined our sample size and all manipulations in the study, and we follow JARS (Kazak, 2018). Here we reported only the data exclusions and measures relevant for the conducted analyses. All data exclusions and measures are described in the study protocol of the OPTIMA trial (Kopf-Beck et al., 2020), of which this study is a sub-study. The dataset generated and analyzed for this study contains clinical data and is not publicly available due to the protection of participants' rights to privacy and data protection but is available from the corresponding author on reasonable request. Materials and analysis code for this study are available by emailing the corresponding author. This study's design, hypotheses and analysis plan was preregistered before the end of the data collection and before analyses were undertaken; see osf.io/9fuhn. Data was analyzed using R, version 4.2.2 (R Core Team, 2020).

Results

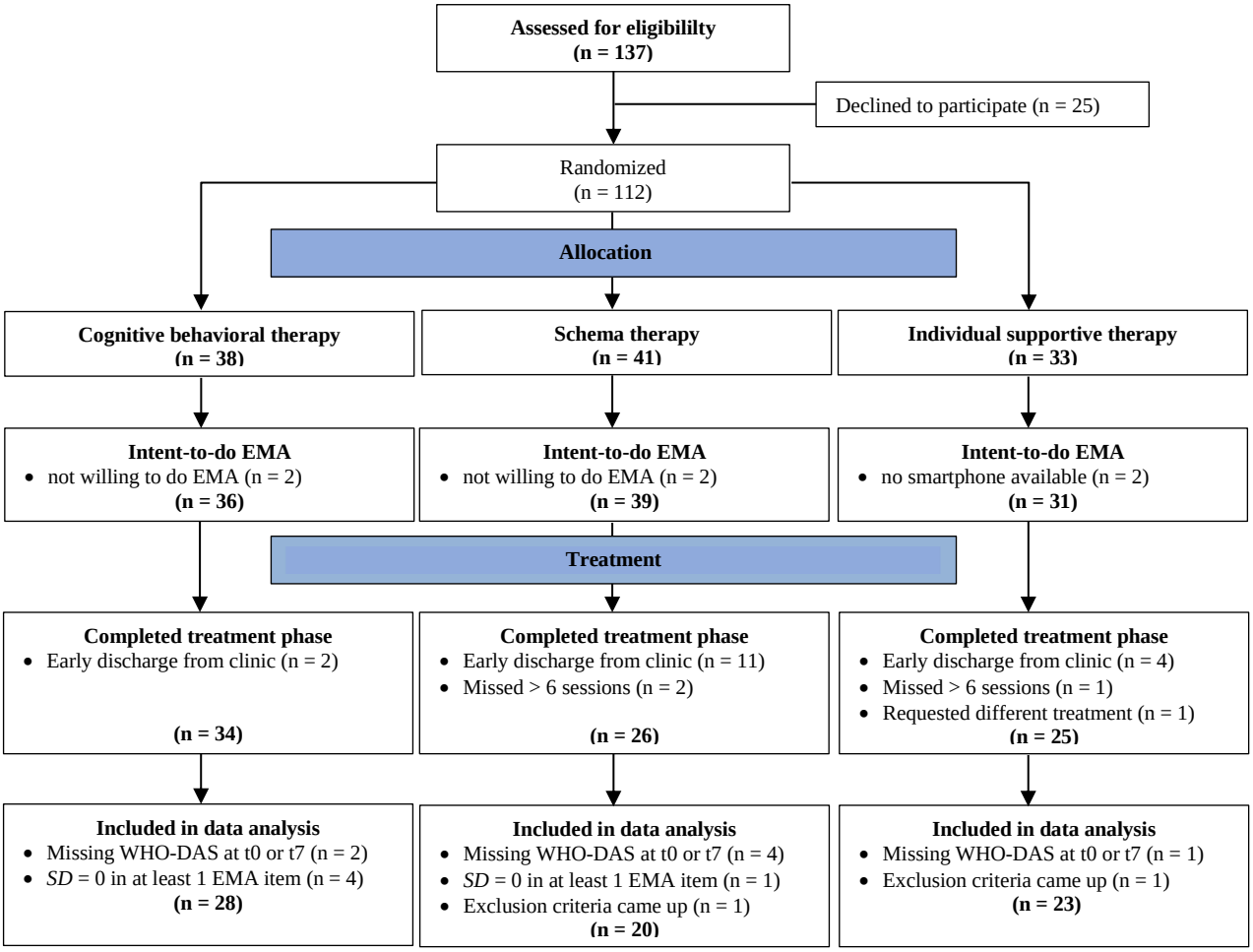
Sample description

The study was conducted from August 2019 to December 2020. Initially, 137 patients were recruited and assessed for eligibility, of which 106 met inclusion criteria, expressed willingness to participate in the EMA, and were randomly assigned to the intervention trial (ST, 39, CBT, 36, IST, 31). Detailed information is presented in the CONSORT flow chart in

Figure 2.1. Survival analyses (for details see Figure A.1 in the Supplementary Material) revealed that patients in CBT exhibited a significantly lower risk of dropping out during the intervention phase compared to those in ST (Cox regression: ß = -1.70, $p = 0.02$; relative risk in CBT= 0.17, 95% confidence interval [CI] = 0.04 – 0.75). IST did not significantly differ from ST ($p = 0.2$) or CBT ($p = 0.66$) regarding dropout risk. Reasons for drop out included early discharge from the clinic before completion of the intervention trial (17), missing more than six intervention sessions (3), and request for a different intervention (1). Before analyses, we further excluded the data of 14 patients due to ineligible diagnoses that surfaced during the study (2), patients that missed one of the clinical interviews (7) or lack of conscientious completion of the EMA (5), which was defined as a standard deviation of zero in at least one EMA item during the intervention trial. Ultimately, data of 71 patients (CBT, 28, ST, 20, IST, 23) were included in the statistical analyses.

Figure 2.1

Data Exclusion Flow Diagram



Note. EMA: Ecological Momentary Assessment; WHO-DAS: World Health Organization Disability Assessment Schedule.

Descriptives of the study sample are presented in Table 2.1 and Table A.3 in the Supplementary Material. At baseline, patients exhibited severe levels of depression on average (BDI-II: $M = 32.57$, $SD = 8.4$). As shown in Table 2.1, patients in the intervention conditions exhibited significant differences in baseline depression and gender. Specifically, holm-adjusted post-hoc t-tests between conditions revealed a significant difference between CBT and the two comparing intervention conditions IST and ST in baseline BDI-II (CBT: $M = 29$, $SD = 7.35$; IST: $M = 35.83$, $SD = 8.62$; ST: $M = 33.8$, $SD = 7.89$; CBT – IST: $p = 0.004$; CBT – ST: $p = 0.037$). Consequently, depression severity at baseline and gender were included as additional covariates in the analyses.

In the analysis sample, patients responded to a mean of 57.77% ($SD = 25.31\%$) of the EMA prompts (Table 2.1). We found significant differences in the distribution of patients' response rates over the intervention weeks ($F(7, 560) = 4.72$, $p < 0.001$), displayed in Figure A.2 in the Supplementary Material. Note that patients had in mean 0.69 (min = 0, max = 11) prompts with missing items. EMA prompts with missing items (N = 49) were filtered out of the analysis data set. The EMA feasibility questionnaire underscored the good acceptance of the EMA among patients, with 92.73% expressing a liking for our app – rating it as very good, good or reasonably good.

Table 2.1

Descriptive Statistics of the Treatment Arms

			Treatment								
	Total (N=71)		ST (n=20)		CBT (n=28)		IST (n=23)				
Characteristic	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**t or Chi²**	**df**	***p***
Gender (female)	39	54.93	15	75.00	10	35.71	14	60.87	7.65	2	0.02
Nationality (german)	60	84.51	16	80.00	25	89.29	19	82.61	0.77	2	0.68
School graduation (Qualification for University)	42	59.15	11	55.00	19	67.86	12	52.17	1.80	2	0.41
Income									5.74	2	0.06
Low income (<1500 EUR)	29	40.85	6	30.00	15	53.57	8	34.78			
Middle income (1500 - 4000 EUR)	27	38.03	10	50.00	6	21.43	11	47.83			
High income (>4000 EUR)	11	15.49	3	15.00	7	25.00	1	4.35			
not specified	4	5.63	0	0.00	0	0.00	3	13.04			
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	**t or Chi²**	**df**	***p***
Age (years)	40.96	12.17	41.35	11.44	38.54	13.65	43.57	10.71	2.07	2	0.36
Response Rates											
EMA	57.77	25.31	57.92	27.22	62.56	24.54	51.81	24.34	2.59	2	0.27
BDI	92.08	13.57	92.50	13.08	93.75	11.02	89.67	16.71	1.03	2	0.60
PTQ	90.49	14.86	91.88	13	91.96	14.52	87.50	16.85	2.33	2	0.31
Baseline Symptoms											
BDI	32.56	8.40	33.80	7.98	29	7.35	35.83	8.62	4.99	2	0.01
PTQ	40.44	11.19	41	12.44	36.54	11.96	44.70	7.11	6.11	2	0.05
WHODAS	2.87	0.65	2.92	0.56	2.67	0.67	3.08	0.67	2.61	2	0.08

Note. ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; BDI-II; Beck's Depression Inventory II; PTQ: Perseverative Thinking Questionnaire; WHO-DAS: World Health Organization Disability Assessment Schedule.

Hypothesis 1

To test whether change in depressive symptoms or RNT assessed with EMA is associated with larger intervention effects than when assessed with WQA, we ran MLMs separately for depressive symptoms or RNT measured with EMA or WQA and calculated standardized effect sizes. For each MLM we compared a simple model including only time

and intervention condition as predictors with a complex model including gender, baseline depression, response rate and concomitant care as covariates. We compared the simple and the complex models with model comparison analyses (Table A.4 in the Supplementary Material). As the complex models explained our data not significantly better than the simple models and no covariates were significant predictors, we report the results of the simple models and also based further analyses on them (Table 2.2). In the MLM analyzing EMA-assessed depressive symptoms, and the MLMs for the two WQA scores (RNT and depressive symptoms), we observed significant time predictors, indicating substantial score changes from baseline to the end of the intervention. Moreover, we identified a significant intervention effect, denoted by a significant time*condition interaction, for the EMA-assessed RNT when comparing ST with both reference intervention conditions, IST and CBT.

Table 2.2

Multi-Level Models of Depression and RNT measured by EMA versus WQA among Patients randomly assigned to ST, IST or CBT, with ST as Reference Group

DV/predictors	**Estimates**	**SE**	**t (*p*)**	**95% CI**
EMA – Depressive symptoms				
Intercept	0.35	0.11	3.06 (0.002)	[0.13 - 0.57]
Condition (CBT – ST)	-0.26	0.15	-1.77 (0.082)	[-0.56 - 0.03]
Condition (IST – ST)	-0.12	0.16	-0.75 (0.459)	[-0.43 - 0.2]
Time	<0.01	<0.01	-3.03 (0.002)	[-0.01 - <0.00]
Time x Condition (CBT – ST)	<0.01	<0.01	1.78 (0.076)	[<0.00 - 0.01]
Time x Condition (IST – ST)	<0.01	<0.01	0.78 (0.437)	[<0.00 - 0.01]
WQA – Depressive symptoms				
Intercept	1.08	0.14	7.78 (<0.001)	[0.81 - 1.35]
Condition (CBT – ST)	-0.14	0.18	-0.75 (0.454)	[-0.5 - 0.22]
Condition (IST – ST)	-0.04	0.19	-0.23 (0.818)	[-0.42 - 0.33]
Time	-0.32	0.04	-8.41 (<0.001)	[-0.39 - -0.24]
Time x Condition (CBT – ST)	0.05	0.05	0.99 (0.32)	[-0.05 - 0.15]
Time x Condition (IST – ST)	0.01	0.05	0.21 (0.836)	[-0.09 - 0.11]
EMA – RNT				
Intercept	0.52	0.10	5.15 (<0.001)	[0.32 - 0.71]
Condition (CBT – ST)	-0.47	0.13	-3.59 (0.001)	[-0.73 - -0.21]
Condition (IST – ST)	-0.33	0.14	-2.37 (0.021)	[-0.6 - -0.05]
Time	-0.01	<0.01	-5.33 (<0.001)	[-0.01 - <0.01]
Time x Condition (CBT – ST)	0.01	<0.01	3.75 (<0.001)	[<0.01 - 0.01]
Time x Condition (IST – ST)	<0.01	<0.01	2.46 (0.014)	[<0.01 - 0.01]
WQA – RNT				
Intercept	0.84	0.18	4.64 (<0.001)	[0.48 - 1.19]
Condition (CBT – ST)	-0.12	0.24	-0.5 (0.62)	[-0.59 - 0.35]
Condition (IST – ST)	-0.21	0.25	-0.87 (0.387)	[-0.7 - 0.27]
Time	-0.25	0.05	-4.88 (<0.001)	[-0.34 - -0.15]
Time x Condition (CBT – ST)	0.04	0.07	0.62 (0.538)	[-0.09 - 0.17]
Time x Condition (IST – ST)	0.06	0.07	0.8 (0.425)	[-0.08 - 0.19]

Note. EMA: Ecological Momentary Assessment; WQA: Weekly questionnaire assessment; ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; RNT: Repetitive Negative Thinking; Sample size in all models is 71.

Furthermore, when looking at the standardized effect sizes between ST and the reference interventions, the Cohen's d effect sizes were higher, and the NNTs (to detect an

additional responder in ST versus IST and CBT) were lower for EMA-assessed than WQA-assessed RNT (Table 2.3).

Table 2.3

Standardized Effect Sizes (Cohen's d and NNT) of Depression and RNT measured by EMA versus WQA among patients randomly assigned to ST, IST or CBT

Variables/ group comparisons	**Cohen's d**	**NNT**
EMA – Depressive symptoms		
IST - ST	0.26	10
CBT - ST	0.62	4
WQA – Depressive symptoms		
IST - ST	0.1	24
CBT - ST	0.47	5
EMA - RNT		
IST - ST	0.87	3
CBT - ST	1.46	2
WQA - RNT		
IST - ST	0.35	7
CBT - ST	0.26	10

Note. EMA: Ecological Momentary Assessment; WQA: Weekly questionnaire assessment; ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; RNT: Repetitive Negative Thinking; NNT: Number-Needed-to-Treat; Sample size in all analyses is 71.

Hypothesis 2

Multiple regression analyses were conducted to test whether change in depressive symptoms or RNT assessed with EMA predicts change in global functioning measured by clinician ratings more strongly than the same predictors assessed with WQA (Table 2.4). We ran four separate models, one for each slope of EMA- or WQA-assessed depressive symptoms or RNT as the focal predictor. Additionally, we performed two combined models (one for depressive symptoms and one for RNT) including both assessment techniques as predictors (EMA and WQA). The separate models revealed that both EMA-assessed and WQA-assessed slopes of depressive symptoms, along with the WQA-assessed slope of RNT, significantly predicted post-intervention global functioning, while the EMA-assessed slope of RNT stayed hovered just below the significance threshold ($p = 0.051$). In the combined

models, none of the EMA slopes significantly predicted our dependent variable, whereas both WQA slopes did. We conducted model comparison analyses to test the combined models including both predictors (EMA and WQA) against the separate WQA models. The AIC and BIC values of the simple models were slightly lower than those of the complex models (AIC/BIC for simple models: Depressive Symptoms – 83.29/92.35, RNT – 85.45/94.5; AIC/BIC for complex models: Depressive Symptoms – 83.1/94.41, RNT – 87.43/98.74), indicating no significant improvement in model fit with EMA slopes as additional predictors.

Table 2.4

Multiple Regression Analyses predicting Global Functioning (GF) measured with Clinical Interview (WHO-DAS) after seven Weeks of Treatment based on Baseline Global Functioning and the Slope of Depression and RNT measured with EMA versus WQA

DV/predictors	**Estimates**	**SE**	**t (p)**	**95% CI**	**F statistic**	**R2**
EMA – Depressive symptoms						
Intercept	0.55	0.24	2.34 (0.022)	[0.08 - 1.02]	F(2.68) = 30.52	0.47
GFpre	0.60	0.08	7.46 (<0.001)	[0.44 - 0.76]	(<0.001)	
Slope (EMA-Dep)	26.13	9.32	2.8 (0.007)	[7.54 - 44.72]		
WQA – Depressive symptoms						
Intercept	1.07	0.27	3.99 (<0.001)	[0.54 - 1.61]	F(2.68) = 35.92	0.51
GFpre	0.58	0.08	7.54 (<0.001)	[0.43 - 0.73]	(<0.001)	
Slope (WQA-Dep)	1.80	0.48	3.77 (<0.001)	[0.85 - 2.76]		
EMA & WQA – Depressive symptoms						
Intercept	0.99	0.27	3.61 (0.001)	[0.44 - 1.53]	F(3.67) = 25.04	0.53
GFpre	0.59	0.08	7.69 (<0.001)	[0.44 - 0.74]	(<0.001)	
Slope (EMA-Dep)	14.28	9.83	1.45 (0.151)	[-5.34 - 33.89]		
Slope (WQA-Dep)	1.48	0.53	2.81 (0.006)	[0.43 - 2.53]		
EMA – RNT						
Intercept	0.57	0.24	2.33 (0.023)	[0.08 - 1.05]	F(2.68) = 27.18	0.44
GFpre	0.59	0.08	7.19 (<0.001)	[0.43 - 0.76]	(<0.001)	
Slope (EMA-RNT)	21.48	10.81	1.99 (0.051)	[-0.09 - 43.05]		
WQA – RNT						
Intercept	0.86	0.25	3.45 (0.001)	[0.36 - 1.36]	F(2.68) = 33.83	0.5
GFpre	0.55	0.08	6.95 (<0.001)	[0.39 - 0.71]	(<0.001)	
Slope (WQA-RNT)	1.09	0.32	3.43 (0.001)	[0.46 - 1.72]		
EMA & WQA – RNT						
Intercept	0.86	0.26	3.34 (0.001)	[0.35 - 1.37]	F(3.67) = 22.24	0.5
GFpre	0.55	0.08	6.85 (<0.001)	[0.39 - 0.71]	(<0.001)	
Slope (EMA-RNT)	1.83	12.64	0.15 (0.885)	[-23.4 - 27.06]		
Slope (WQA-RNT)	1.06	0.39	2.7 (0.009)	[0.28 - 1.84]		

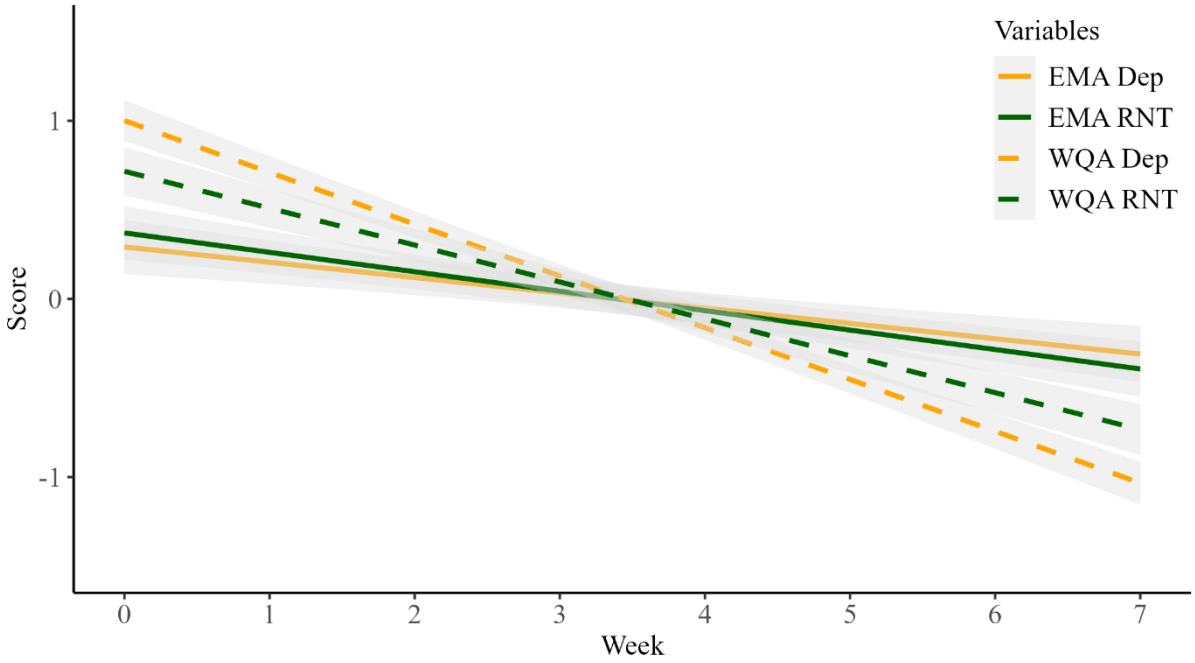
Note. RNT: Repetitive Negative Thinking; Dep: Depression; GF: Global Functioning; EMA: Ecological Momentary Assessment; WQA: Weekly Questionnaire Assessment; All models have two degrees of freedom. Sample size in all models is 71.

Additional analysis

Figure 2.2 illustrates the linear modeling of the reduction in patients' depressive symptoms and RNT, as assessed by EMA and WQA, throughout the seven-week intervention period (please note that the scores are person-mean centered and z-standardized).

Figure 2.2

Comparison of EMA versus WQA Data in the Time Series



Note. The figure illustrates the linear modeling of the reduction in patients' depressive symptoms (orange) and RNT (green) measured with EMA (solid lines) versus WQA (dashed lines) during the seven-week intervention period. All scores are z-standardized and person mean centered. EMA Scores are aggregated to weekly means and 95% confidence intervals are shown (light grey). EMA: Ecological Momentary Assessment; WQA: Weekly Questionnaire Assessment; RNT: Repetitive Negative Thinking.

To dive deeper into the distinctions between EMA and WQA, we conducted two further MLMs (one for depressive symptoms and one for RNT), including 'assessment technique' (EMA vs. WQA) as a predictor (Table 2.5). In both models, we found significant time*assessment technique interactions, revealing significant differences in time effects (slopes) between EMA and WQA. Given that our EMA and WQA scores predominantly comprised different items, we also ran corresponding MLMs at the level of three individual

RNT items that were identical between the EMA and WQA scores of RNT. At this item level, the results were similar to those obtained with the full-item score of RNT.

Table 2.5

Multi-Level Models of Depression and RNT measured by EMA versus WQA with Assessment Technique (EMA versus WQA) as Predictor

DV/predictors	**Estimates**	**SE**	**t (or z)**	**95% CI**
Depressive symptoms				
Intercept	0.29	0.09	3.37 (0.001)	[0.12 - 0.46]
Time	-0.09	0.02	-3.57 (<0.001)	[-0.13 - -0.04]
Assessment Technique	0.72	0.09	8.28 (<0.001)	[0.55 - 0.89]
Assessment Technique*Time	-0.21	0.02	-9.74 (<0.001)	[-0.25 - -0.16]
RNT				
Intercept	0.37	0.10	3.82 (<0.001)	[0.18 - 0.55]
Time	-0.11	0.03	-3.99 (<0.001)	[-0.16 - -0.06]
Assessment Technique	0.36	0.09	3.92 (<0.001)	[0.18 - 0.53]
Assessment Technique*Time	-0.10	0.02	-4.65 (<0.001)	[-0.15 - -0.06]
RNT Item 'repetitiveness'				
Intercept	0.38	0.09	4.3 (<0.001)	[0.21 - 0.56]
Time	-0.11	0.02	-4.62 (<0.001)	[-0.16 - -0.07]
Assessment Technique	0.26	0.10	2.69 (0.007)	[0.07 - 0.45]
Assessment Technique*Time	-0.08	0.02	-3.21 (0.001)	[-0.12 - -0.03]
RNT Item 'intrusiveness'				
Intercept	0.35	0.09	3.8 (<0.001)	[0.17 - 0.53]
Time	-0.10	0.03	-4.08 (<0.001)	[-0.15 - -0.05]
Assessment Technique	0.20	0.10	2.06 (0.04)	[0.01 - 0.39]
Assessment Technique*Time	-0.06	0.02	-2.46 (0.014)	[-0.11 - -0.01]
RNT Item 'controllability'				
Intercept	0.33	0.09	3.57 (<0.001)	[0.15 - 0.51]
Time	-0.10	0.03	-3.87 (<0.001)	[-0.15 - -0.05]
Assessment Technique	0.23	0.10	2.32 (0.021)	[0.04 - 0.42]
Assessment Technique*Time	-0.06	0.02	-2.71 (0.007)	[-0.11 - -0.02]

Note. EMA: Ecological Momentary Assessment; WQA: Weekly Questionnaire Assessment; Dep: Depression; RNT: Repetitive Negative Thinking; Sample size in all models is 71.

Discussion

This study compared two different techniques, EMA and WQA, to measure change in depressive symptoms and RNT within a clinical trial. The primary objectives were twofold: a)

to determine their sensitivity in detecting intervention effects between conditions, and b) to assess their validity regarding the prediction of clinician-rated global functioning. The study underscores the feasibility of continuous EMA throughout the whole course of a clinical trial in moderately to severely depressed patients. To the best of our knowledge, it is the first clinical trial studying this technique in comparison to a questionnaire setup that is usual in clinical trials: weekly assessment of full questionnaires.

Intervention effects

In a study investigating the effects of mindfulness therapy on mindfulness, depression, and anxiety, Moore et al. (2016) identified higher intervention effects for EMA-assessed mindfulness and depression compared to questionnaire-assessed measures. This indicates a heightened sensitivity of EMA in detecting significant intervention effects. Our results support this assumption. Based on these findings by Moore et al. (2016) we aimed to examine whether changes in depressive symptoms and RNT assessed with EMA would result in larger intervention effects than those assessed with WQA. Indeed, our analysis revealed that EMA-assessed RNT is associated with significant intervention effects, accompanied by higher effect sizes and lower NNTs compared to the WQA variable. Notably, contrary to our initial hypothesis, we observed this effect only for RNT, while for change in depressive symptom we found no significant intervention effects, regardless of whether they were assessed with EMA or WQA. However, this does not undermine the underlying assumption of EMA's heightened sensitivity in detecting such effects. Interestingly, our findings for depressive symptoms align with the results of the OPTIMA study (Kopf-Beck et al., 2024), of which our study is a sub-study. Including a larger population of 292 patients, even the OPTIMA study (Kopf-Beck et al., 2024) found no significant intervention effects on depressive symptoms between ST and the two comparing interventions, CBT and IST. This supports the notion that ST is indeed comparably effective in reducing depressive symptoms when compared to CBT and the active control condition IST and serves as a plausible explanation for why not even EMA detected intervention effects between these conditions for changes in depressive symptoms in our study.

Prediction of change in global functioning

Studying the effects of a 6-week antidepresssant trial, Targum et al. (2021) discovered that continious EMA of changes in depression significantly predicted clinician-rated outcomes. To corroborate these findings, we analyzed the validity of EMA and WQA in

predicting changes in global functioning. We examined them in separate models (EMA or WQA) and in combined models incorporating both predictors (EMA and WQA). Changes in depressive symptoms and RNT assessed with WQA significantly predicted changes in global functioning in both the separate and combined models. In contrast, the results for EMA were mixed. Specifically, changes in depressive symptoms assessed with EMA emerged as a significant predictor in the separate model, but not in the combined model, and EMA-assessed changes in RNT failed to significantly predict change in global functioning in either the separate or combined model. The results are consistent with those of Targum et al. (2021), in that EMA-assessed changes in depression significantly predict clinician ratings. Nevertheless, our hypothesis, asserting that EMA-assessed changes in depressive symptoms and RNT could be stronger predictors of clinician-rated changes in global functioning, was not supported by the results. A plausible explanation for our findings could be the shared time reference of questionnaires and clinician ratings, as both, unlike EMA, rely on retrospective recall.

Additional Analysis

In light of these results, we conducted additional analyses and discovered significant differences in the time effects (slopes) of EMA and WQA. The WQA scores exhibited a steeper decrease over time, with more extreme scores in both directions - higher at baseline and lower post- intervention - compared to the EMA scores. The effect was evident for changes in RNT and depressive symptoms, as well as at the level of individual RNT items that were identical between EMA and WQA. This observation indicates that the effect is not solely attributable to differences in item selection between the EMA and WQA scores of depressive symptoms and RNT. A plausible explanation for these findings could be the presumed recall bias associated with questionnaires: While in the past it was predominantly assumed that memories of depressed patients are negatively biased (Greenberg & Beck, 1989), more recent literature suggests that memories exaggerate reality in both negative and positive valence. This bias is amplified in depressed patients - stronger in terms of negative than positive bias, but still in both directions (Ben-Zeev et al., 2009). Theoretically, this effect could lead to overestimations of time effects in depression trials when relying on retrospective questionnaires. As depressed individuals are more susceptible to recall bias than healthy individuals, it’s plausible that retrospectively they may overestimate the severity of their symptoms at the beginning of a therapy, while when they are less depressed at the end of a therapy, their retrospective self-reports might be more realistic or even positively biased. Our study supports this assumption. Nevertheless, alternative explanations for these findings

cannot be ruled out. For instance, it's plausible that EMA might exhibit smaller change amplitudes over time due to anchoring effects. When individuals are frequently rating the same items, as with EMA, previous responses may serve as reference points, influencing subsequent answers. Consequently, patients' EMA ratings may exhibit high interdependence, whereas their questionnaire ratings, collected at one-week intervals and containing a larger number of items, may elicit more independent responses.

Limitations

This study has several limitations. To our knowledge, this was the first study comparing EMA and WQA monitoring of depression-related constructs using different items in terms of quantity and wording. While this approach increases the external validity of our results, it also lowers the comparability between our EMA and WQA outcomes compared to previous studies. A 1:1 transfer of questionnaires into an EMA setting is difficult, as questionnaires usually have a substantial number of items. Administering such a high item count on a daily basis or multiple times per day could overly burden patients. Previous studies have therefore compared questionnaire subscales between EMA and questionnaires, which deviates from the standard of clinical trials, where full questionnaires are usually administered. In our additional analysis however, we could demonstrate that significant differences between EMA and WQA slopes persist even at the level of identical items.

Additionally, it is essential to consider that clinical interviews might not serve as the best benchmark for comparing the predictive validity of EMA and WQA. Both, questionnaires, and clinical interviews collect symptoms retrospectively, carrying the risk of recall biases. Even though clinical interviews are third-party ratings conducted by trained personnel, which should reduce this risk (Malhi et al., 2017), the retrospectivity of both assessment methods may explain why questionnaires predict the results of clinical interviews stronger than EMA.

Furthermore, our sample exhibited variations across the three intervention conditions in terms of demographical and clinical aspects. Despite our efforts to control for these variables in our analyses, the inherent heterogeneity of the sample, coupled with its modest size and the fact that psychotherapy was an important, but just a part of the comprehensive treatment program, compounds the interpretation of effects between conditions. Additionally, the monocentric study design, in which most patients were from the area of munich, Germany, limits the generalizability of our study due to selection bias.

Conclusion

Aligned with prior research, our study establishes significant associations among clinical interview, questionnaires and EMA ratings of change in depression. Despite substantial evidence supporting EMA's ability to detect higher effect sizes between intervention conditions than questionnaires, our findings, unlike Moore et al. (2016), suggest a more nuanced perspective. The assumption that EMA is simply more sensitive to change falters, as our study indicates smaller change amplitudes in EMA compared to questionnaires. Therefore, an alternative perspective emerges, suggesting that EMA provides more accurate estimates, enhancing statistical power and resulting in clearer effects, as indicated by a recent simulation study (Schuster et al., 2020). Sampling and/or memory biases, which are inherent in depressed patients, may undergo changes during therapy, thereby systematically influencing the repeated assessments of retrospective questionnaires. Especially, when investigating intervention effects in modest sample sizes, coupled with active control conditions, which lower the expected effect size differences, this systematic bias may hinder the detection of significant effects. Therefore, we assume it is not EMA's heightened sensitivity to symptom change but its lower levels of sampling and memory biases making it more sensitive to intervention effects.

Nevertheless, our study prompts questions about what each instrument truely measures and which is most valid for monitoring depression-related change. Beyond recall bias, factors like current mood or expectations introduce biases into retrospective self-reports. While clinical interviews aim to mitigate such biases, it is uncertain if they are free of them and their nature of external ratings introduce other potential biases, such as influences of the rater on the social desirability of the interviewee. It can be assumed that the momentary nature of EMA inbedded in the natural environment of the patient avoids these biases, but however, beeing asked the same question several times per day introduces risks of low conscientiousness or anchoring effects. Another perspective is to consider questionnaires more as a trait-measure and EMA more as a state-measure of depression. The EMA in our study sampled momentary states ("right now" states), whereas the questionnaires captured an aggregated subjective measure of depression over time (e.g., two weeks in case of BDI-II). Complementing questionnaires with EMA could therefore become a progressive approach that allows the investigation of three distinct aspects: the momentary change in depressive symptoms, patients' perceived change, and the discrepancy between the two as a reflection of change in memory bias. In conclusion, all three assessment techniques - EMA, questionnaires

and clinical interviews - pose distinct strengths, but also biases challenging their validity. Therefore, it might be promising to zoom out further and consider not retrospectivity but subjectivity as the fundamental problem of self-ratings. On this point, questionnaires, clinical interviews, and even EMA ratings have their limits, as they are all based on self-reports. Therefore, it might be a promising avenue to develop more creative EMA-approaches that combine self-reports with objectively measured behavioral markers of depression, such as homestay, social avoidance, physical activity and sleep (Angel et al., 2022).

3. Study II:

Early Improvement predicts Treatment Response in Depression: An Ecological Momentary Assessment Study

This chapter is a pre-peer-review, pre-copyedit version of a manuscript submitted to the *Behavior Therapy*.

Tamm, J., Takano, K., Just, L., Ehring, T., Rosenkranz, T., OPTIMA study group & Kopf-Beck, J. (under review) Early Improvement predicts Treatment Response in Depression: An Ecological Momentary Assessment Study. Manuscript submitted to *Behavior Therapy*.

Abstract

Objective: Predicting treatment response through early improvement can reduce patients' time in ineffective treatments before considering alternatives. However, for psychological interventions, there is no consensus on what time window and improvement rate early in the treatment is the most informative for distinguishing treatment responders from non-responders. This study investigated these aspects and compared Weekly Questionnaire Assessments (WQA) and Ecological Momentary Assessment (EMA) regarding their power to predict treatment response through early improvement. **Method**: Fifty-two depressed patients were randomly assigned to one of three seven-week psychological interventions (two individual and two group sessions per week). Early improvement was assessed three times daily with EMA and weekly with questionnaires (BDI-II). Linear Regression Models and Receiver Operating Characteristic Analyses were conducted to predict treatment response (BDI-II improvement from pre- to post-intervention ≥ 50%) and ratios of true negative/false negative predictions were calculated to explore the predictive value of different early improvement definitions: 10%, 20%, 30% or 40% improvement after one, two, three or four treatment weeks. **Results**: Both, EMA and WQA significantly predicted responder status after three weeks with AUC values of 73% (EMA) and 77% (WQA). A WQA-assessed 10% improvement after four weeks yielded the highest ratio of true negative/false negative predictions, with a true negative rate of 22% and a false negative rate 0%. **Conclusions**: 10% improvement in depressive symptoms assessed with WQA after three to four weeks of treatment was the best predictor in our study. Further research is needed to validate the results.

Introduction

It is well established that there are several effective psychological interventions for the treatment of depression (Cuijpers, 2015). Besides cognitive behavioral therapy (CBT), interventions such as schema therapy (ST) have proven to effectively reduce depression (Kopf-Beck et al., 2024). However, despite these advancements, meta-analytic findings show that almost 60% of depressed patients do not adequately respond to these treatments (Cuijpers, Karyotaki et al., 2021), which is commonly defined as a reduction of depressive symptoms from baseline to the end of treatment by at least 50% (Rush et al., 2006). Combined treatments of psychotherapy and pharmacotherapy show better outcomes (Cuijpers, Quero et al., 2021), but even there is room for improvement. This means, although different interventions yield similar average effects, treatment responses are highly variable on the individual level. Assuming that non-responders of one treatment might respond to other interventions (Gloster et al., 2020; McKay et al., 2010), it is essential to develop decision rules that can guide clinicians in tailoring treatments to individual patients.

Besides efforts to develop personalized interventions by allocating patients to optimal treatments from the outset of therapy (DeRubeis et al., 2014; Huibers et al., 2015), it is essential to investigate whether and how non-response to an ongoing treatment can be predicted early-on. Treatment prediction based on early improvement monitoring is a promising approach, as it may reduce the amount of time patients spend in ineffective treatment before considering alternatives. In this way, early treatment prediction may lead to better clinical outcomes (Schaffer et al., 2013) while saving scarce resources of the mental health care system (Richards, 2012). Additionally, it could be used in innovative trial designs such as the ‘leapfrog’ method (Blackwell et al., 2019), which involves rapidly testing and modifying treatments based on early indicators of effectiveness.

Indeed, there is robust and replicated evidence that early improvement in therapy is a reliable prognostic indicator for treatment outcome in depression (Rubel et al., 2015). For pharmacological treatment, the absence of early improvement within the initial two weeks of therapy has been identified as a strong indicator of treatment non-response (Szegedi et al., 2009) and the absence of early improvement within the first four weeks has been established as a guideline for clinical decisions regarding medication change (Gautam et al., 2017). For psychological treatments however, such guidelines are lacking. While compelling evidence suggests that early improvements can predict treatment outcomes in psychological

interventions as well, studies exhibit significant heterogeneity, lacking a standardized definition for predictive early improvement in depression (Beard & Delgadillo, 2019; Li et al., 2023). Importantly, earlier studies have used very different time windows classified as 'early' and different definitions for the rate of symptomatic change classified as 'improvement'. For example, 'early' time windows in earlier research could encompass two, four, six or eight weeks of treatment, and early improvement has been defined either using cut-offs for symptom reduction (e.g., >25%, Gois et al., 2014), the achievement of reliable or clinically significant improvement (Rubel et al., 2015), or the occurrence of sudden gains (Hunnicutt-Ferguson et al., 2012).

Moreover, to our knowledge, except for one online intervention study (Schibbye et al., 2014), none of the existing studies has used Ecological Momentary Assessment (EMA) as an alternative technique to monitor early improvement. However, EMA might be better suited for this purpose. It is well known that depressive symptoms naturally fluctuate during the day and from day to day (Chen et al., 2022; Takano & Tanno, 2011) and that the human recall of past affective experiences is biased, especially in depression (Colombo et al., 2020). Consequently, it is questionable whether a single retrospective point assessment can accurately and reliably capture the true symptom severity patients experience over a past week or weeks. EMA may therefore assess change in depression more reliably, especially when observing short time windows (Moore et al., 2016; Tamm et al., 2024).

Thus, we assessed early improvement in depression comparing EMA and Weekly Questionnaire Assessment (WQA) and investigated their predictive value for identifying treatment responders and non-responders across three different seven-week psychological interventions. In addition, we compared four different time windows (one, two, three or four weeks after treatment initiation) and four different definition of early improvements operationalized as symptomatic change rates (minimum improvement of 10%, 20%, 30% or 40%). In this way, we addressed the following research questions: (1a) At which time point after treatment initiation does early improvement in depressive symptoms significantly predict treatment response? (1b) Do both, WQA and EMA of early improvement significantly predict treatment response at these time points? (2) How predictive are different definitions of early improvement in terms of the defined time window and symptom cutoff, as assessed by EMA versus WQA? While other authors aimed to maximize the sum of sensitivity (rate of correctly identified responders, i.e., true positive rate) and specificity (rate of correctly identified non-responders, i.e., true negative rate) using the youden-index (Crits-Christoph et al., 2001), we

decided to evaluate our predictors based on a reverse version of the negative likelihood ratio (Bolin & Lam, 2013). This approach takes the following critical implications for clinical decision-making into account: We defined that the primary goal of treatment prediction through early improvement would be to identify a high number of non-responders to consider alternative interventions for such early in treatment. The primary goal is therefore to achieve a high true negative rate. At the same time, it is particularly important to avoid the disruption of effective treatments, which means targeting a low false negative rate. Although erroneously recommending the continuation of a treatment to a non-responder would also be unbeneficial, we deemed these scenarios less critical since they reflect the current clinical reality without decision rules anyway. We therefore decided that our best definition of early improvement would be the highest ratio between the true negative rate (i.e., specificity) and the false negative rate (i.e., 1-sensitivity rate), which is a reverse of the negative likelihood ratio described in the literature (Bolin & Lam, 2013).

We hypothesized that (1a) early improvement in depressive symptoms predicts treatment response within the first four treatment weeks, (1b) respectively of whether the early improvement is assessed with EMA or WQA. Research question (2) investigating the predictive value of different definitions of early improvement was kept exploratory. For our analyses we utilized data from the OPTIMA study (Kopf-Beck et al., 2024), a clinical trial investigating the effectiveness of three psychological interventions, cognitive behavioral therapy (CBT), schema therapy (ST), and individual supportive therapy (IST). The study included moderately to severely depressed patients treated in an inpatient or day clinic setting.

Materials and Methods

Participants

The sample size of the study was designed for its primary analysis on treatment effects between intervention conditions assessed by EMA, which is described elsewhere (Tamm et al., 2024). Here, we present a secondary analysis of the data. For the primary analysis, a required sample size around 20 per condition, i.e., $n = 60$ patients in total, was targeted.

The inclusion criteria were aged between 18 and 65 years, a diagnosis of a major depressive disorder, single episode or recurrent, moderate or severe without psychotic symptoms diagnosed by clinical assessment. Additionally, patients had to own a smartphone and agree to the EMA. Participants were compensated according to their EMA response rate.

Drop-outs were defined as enrolled patients of whom incorrect in- or exclusion criteria emerged during the study, who left the clinic before completing the intervention, or missed more than six (22%) intervention sessions. Drop-outs were excluded from analyses due to incomplete data for the calculation of their responder status.

Design and Procedures

The data was gathered at the Max Planck Institute of Psychiatry in Munich, Germany, within the OPTIMA study (Identifier on clinicaltrials.gov: NCT03287362). The OPTIMA study is a monocentric, prospective, rater-blinded, parallel-group, block-randomized clinical trial incorporating repeated measures and three distinct intervention conditions (CBT, ST, and IST). Ethical approval for the study protocol was obtained from the Institutional Ethical Committee of the Faculty of Medicine at LMU Munich (Project number 17–395). Participants provided written informed consent prior to study assessments and randomization. Further details of the study procedures are described in the study protocol (Kopf-Beck et al., 2020).

Interventions

Participants of the OPTIMA study were randomly assigned to CBT, ST, or IST. Each intervention condition lasted seven weeks and comprised two single (50 minutes each) and two group sessions (100 minutes each) per week. All three intervention conditions proved clinical utility in the treatment of depression (Kopf-Beck et al, 2024). Detailed descriptions of the interventions including the intervention manuals are provided in the study protocol and the primary analysis of the OPTIMA trial (Kopf-Beck et al., 2020; Kopf-Beck et al., 2024).

It is important to note that concurrent pharmacotherapy or additional treatments that are common in inpatient or day clinic settings, such as ergotherapy or case management, were not regulated. Decisions on this were at the discretion of the attending psychiatrist but were meticulously documented as control variables. Concomitant care conditions showed no significant effects on treatment differences between intervention conditions in the pimary analysis of the OPTIMA study (Kopf-Beck et al., 2024).

Measures

Details about all measures performed in the OPTIMA study, including a range of questionnaires, clinical interviews, imaging, and tests, are provided in the study protocol (Kopf-Beck et al., 2020). Here, we report only the measures included in our analysis, which encompass two primary outcome variables: A questionnaire and an EMA score assessing

depressive symptoms. An assessment plan is provided in Table B.1 in the Supplementary Material.

EMA

EMA was administered continuously throughout the entire intervention period, starting immediately after patients' enrolment in the study. It involved three prompts per day and was signal-contingent, meaning patients were automatically prompted by their device. During the app onboarding process, patients reported their individual approximate wake-up times. Based on these times, each day was divided into three phases: morning = two hours before to five hours after wake-up time, noon = five to ten hours after wake-up time, evening = ten hours after to two hours before wake-up time. Within each phase, one EMA prompt was emitted with a semi-randomized reminder (signal). Prompts could only be completed within their assigned phase. Average times between patients' responses were calculated (M = 304.8 min (SD = 94.47 min) (morning to noon), M = 339.5 min (SD = 106.51 min) (noon to evening) and M = 809.3 min (SD = 109.63 min) (evening to morning), showing no temporal clustering of responses. To counteract sequence effects, the order of EMA items was randomized across prompts. This protocol, including baseline, resulted in a total of 168 EMA prompts (56 days * 3 prompts per day).

The EMA score of depressive symptoms is a sum score of four EMA items, formulated by JT and JKB in collaboration with clinicians and aligning with diagnostic criteria for major depression (ICD-10). The items encompass three core symptoms of depression (loss of interest, withdrawal, and psychomotor agitation/inhibition) and current mood (the item wordings are provided in Table B.2 in the Supplementary Material). The response scale of the items 'loss of interest', 'withdrawal', and 'psychomotor agitation/inhibition' was two-tiered: Initially, participants responded to a binary *Yes-No* scale. If *Yes* was chosen, participants provided further feedback using a five-point Likert scale, gauging their level of agreement (*not at all, a bit, moderately, considerably, very much*). Conversely, if *No* was selected, the Likert scale was omitted, and the subsequent item followed. The 'mood' item was rated by selecting one of five emojis (*very good, good, moderate, bad, very bad*) that best represented their current mood.

The internal reliability of the EMA total score was assessed within the primary analysis of the data and showed good within-participant reliability and excellent between-participant reliability (for details see Tamm et al., 2024).

Weekly Questionnaire Assessments

The corresponding questionnaire scores of depressive symptoms were assessed weekly with the German version of the Beck Depression Inventory II (BDI-II; Hautzinger et al., 2009), resulting in a total of 8 assessment points, including baseline.

Definition of the Independent Variable Early Improvement

We explored the predictive potential of four distinct time windows and four distinct change rates of early improvement in depressive symptoms: a minimum improvement of 10%, 20%, 30% or 40% by week one, two, three or four after treatment initiation assessed with EMA or WQA. As stated before, the EMA score was derived by aggregating four EMA items, while the WQA score was the BDI-II sum score. The four time windows were selected based on previous studies of early improvement, reviewed by Beard and Delgadillo (2019). Prior research focused on the two-weeks and four-weeks' time windows. We decided against time windows longer than four weeks, as Rubel et. al (2015) have shown that most change in patients' progress occurs by the third treatment session, which corresponds to a treatment duration of less than one week in our study setup. Similar observations derive from pharmacological studies (Schaffer et al., 2013). In recognizing the ecological benefits of 'early' treatment prediction in guiding clinical practice, we expanded our investigation to include weeks one and three, alongside weeks two and four, aiming to shed light on 'how early' early improvements predict psychological treatment outcomes. The symptom cutoffs were chosen based on considerations that they should be smaller than the treatment response (50% improvement) rate, contain similar change rates investigated by other studies (e.g., Gois et al., 2014) and have equal space between each cutoff.

Definition of the Dependent Variable Treatment Response

In line with other studies (Keller, 2003; Rush et al., 2006), treatment response was defined as a reduction of greater than or equal to 50% in the BDI-II total score from baseline to the end of the intervention. For the definition of treatment response, we also considered other types of outcomes, e.g., quality of life. However, we ultimately decided against including them as global mental health indices such as quality of life do not change as markedly within a short duration of psychotherapy (Trivedi et al., 2006).

Statistical Analyses

Patients with insufficient EMA data for calculating early improvement were excluded from analyses. To analyze a change rate in the first treatment week, at least two data points are needed. For analyzing change rates over two, three or four weeks, at least one additional data point per subsequent week is needed. Therefore, patients with fewer than two EMA prompts in the first week or fewer than one EMA prompt in the second, third or fourth week were excluded. For plausibility, we examined the person-level standard deviations for each EMA item throughout the trial period and excluded patients exhibiting a standard deviation of zero in at least one EMA item. For all analyses, we defined significance as $p = 0.05$.

Research question 1a and 1b

We hypothesized that (1a) early improvement in depressive symptoms would predict treatment response within the first four treatment weeks, (1b) regardless of whether early improvement is assessed with EMA or WQA. To test these hypotheses, we conducted a multi-step analysis using linear regression models (LRMs) and logistic regression models. For each participant, we fitted a LRM with time predicting EMA-assessed or WQA-assessed depressive symptoms. The models were estimated for each of the four time-window conditions (i.e., one, two, three, and four weeks) - therefore, we obtained four pairs of individual participants' estimates of the intercept and slope (of time), each for EMA and WQA.

In the second step, an Improvement Rate Score was calculated for each patient and each time-window condition. The Improvement Rate was operationalized as the ratio of symptom change over a given time window (slope*time window) being divided by the baseline symptom level (intercept).

Lastly, we conducted binary logistic regressions to predict responder status (responder vs. non-responder) after seven weeks of treatment through individual baseline BDI-II and the Improvement Rate Score as independent variables.

Research question 2

To determine the 'best' definition of early improvement, considering both the designated time window and the symptom cutoff, we first performed Receiver Operating Characteristic (ROC) analysis for each variable (EMA or WQA) and time window. The ROC analysis is a method commonly used to evaluate the diagnostic performance of a test or model

by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across different threshold values. The Area Under the Curve (AUC) derived from the ROC curve serves as a summary measure of discriminative ability, where higher AUC values indicate better discrimination between responder and non-responder groups (Hanley & McNeil, 1982). To evaluate significant differences between the AUCs of the ROC curves, model comparison analyses were employed using the DeLong's test (DeLong et al., 1988).

Next, we analysed the true negative rate (specificity), false negative rate (1-sensitivity) and the ratio between the two for each predefined combination of variable (EMA or WQA), time window (one, two, three or four weeks) and symptom cutoff (minimum improvement of 10%, 20%, 30% or 40%). In the literature, the inverted form of this ratio ((1-sensitivity)/specificity) is described as the 'negative likelihood ratio (LR-)', displaying the likelihood of a patient testing negative while having a disease (false negative rate) divided by the likelihood of a patient testing negative while not having a disease (true negative rate; Bolin & Lam, 2013). For our research however, a reverse LR- (i.e., specificity/(1-sensitivity)), which we called the TNFN ratio for clarity is better suited and easier to interpret. In our context specificity (the true negative rate) denotes the rate of non-responders correctly identified, while 1-Sensitivity (the false negative rate) denotes the rate of responders falsely classified as non-responders. The TNFN ratio therefore gives us the likelihood of a patient having no early improvement and who does not respond to treatment (true negative rate) divided by the likelihood of a patient having no early improvement but who does respond to treatment (false negative rate). Definied like this, the highest TNFN ratio displays our 'best' predictor, i.e., our 'best' definition of early improvement. To obtain if the TNFN ratio results in more true negative than false negative predictions under the given responder rate in our sample, we also calculated a weighted TNFN ratio (wTNFN ratio) by multiplying the TNFN ratio by the ratio of the non-responder to responder rate: wTNFN ratio (RR) = TNFN ratio*((1-RR)/RR). Note that the responder rate (RR) must lie between zero and one. A higher number of true negative to false negative predictions is indicated by wTNFN > 1.

Finally, to improve the comparability of our results to other studies, we also calculated youden indices (Youden, 1950), which serve as composite measures taking the sum of specificity (true positive rate) and sensitivity (true negative rate) into account (youden index = sensitivity + specificity - 1).

Transparency and Openness

We report how we determined our sample size and all manipulations in the study, and we follow JARS (Kazak, 2018). Here, we present only the pertinent data exclusions and measures relevant to the reported analyses. All data exclusions and measures are described elsewhere (Kopf-Beck et al., 2020; Kopf-Beck et al., 2024). The dataset utilized in this study, comprising clinical data, is not publicly accessible to protect participants' rights to privacy and data protection, but will be provided by the corresponding author upon reasonable request. This also applies to the materials and the analysis code. The study design was preregistered (for details see osf.io/9fuhn). Data was analyzed using R, version 4.2.2 (R Core Team, 2020).

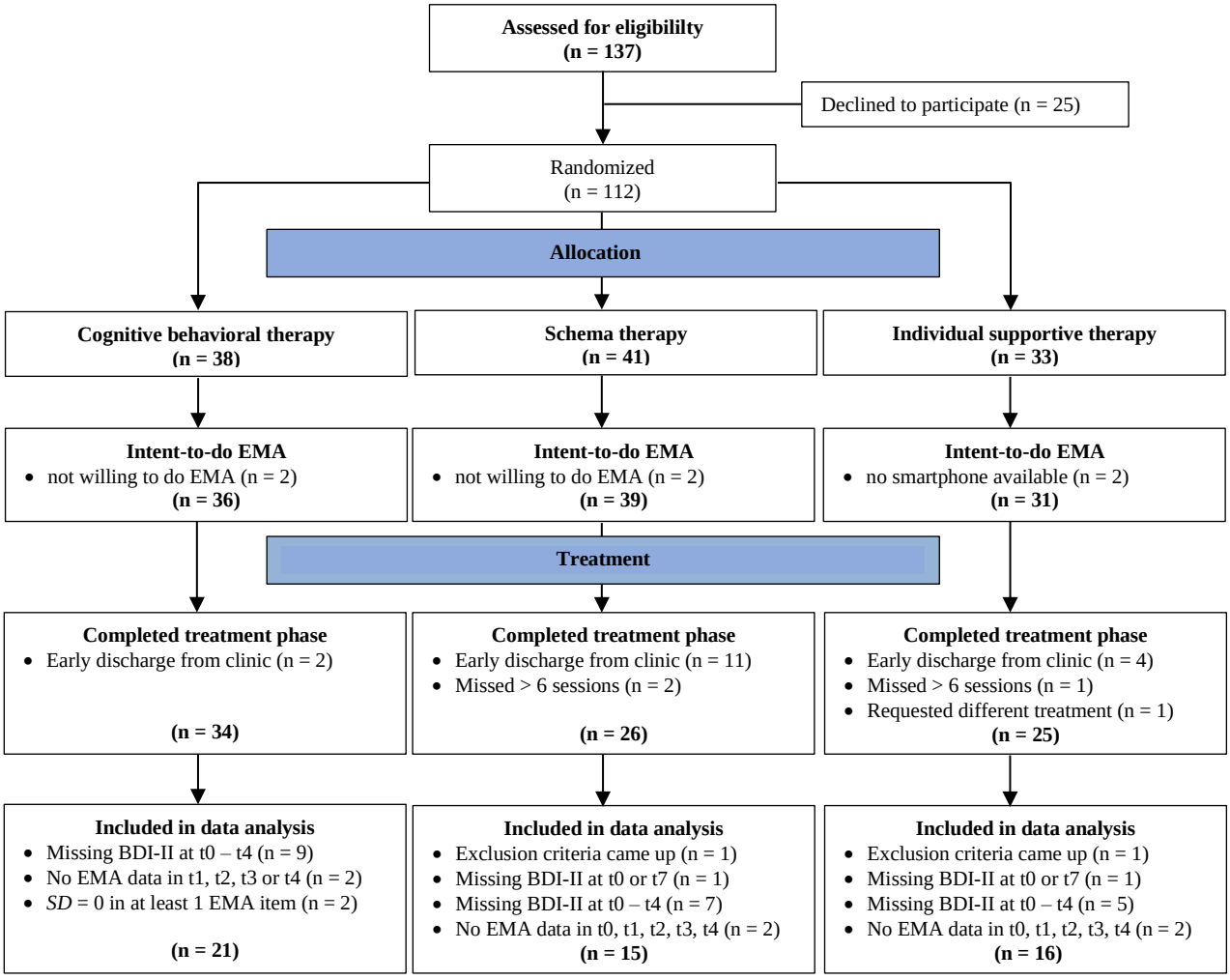
Results

Sample description

Patients were recruited between August 2019 and December 2020. Initial 137 patients were evaluated for eligibility, with 106 meeting the inclusion criteria, expressing willingness to engage in EMA, and being randomly assigned to the intervention conditions (ST: 39, CBT: 36, IST: 31). Details are provided in the CONSORT flow chart in Figure 3.1. Reasons for dropout included early discharge from the clinic prior to completing the intervention trial (17), missing more than six intervention sessions (3), and requesting a different intervention (1). Before conducting analyses, we further excluded data from patients due to ineligible diagnoses identified during the study (2), missing data in the BDI-II baseline or post-intervention assessment (2), missing BDI-II data in the first four intervention weeks (11), missing EMA data in the first four intervention weeks (6), or lack of data integrity in the EMA (2). Finally, the statistical analyses included the data of $N = 52$ patients.

Figure 3.1

Data Exclusion Flow Diagram



Note. EMA: Ecological Momentary Assessment; BDI-II; Beck's Depression Inventory II.

As shown in Table 3.1, patient's age in the analyzed cohort was M = 40.5 years (SD = 11.71), 57.69% were female (n = 30), 84.62% were german (n = 44), and 51.92% had qualifications for university entrance (n = 27). Patients exhibited severe levels of depression on average at baseline (BDI-II: M = 32.58, SD = 8.5). Patients EMA response rate in the analyzed sample was M = 64.67% (SD = 23.35%). Further descriptive statistics such as comorbidities are provided in Table B.3 in the Supplementary Material. Distributions of the EMA response rates across intervention weeks are provided in Figure B.1 in the Supplementary Material.

Table 3.1

Descriptive Statistics of the Sample

			Responder Status						
	Total		Responder		Non-Responder				
Characteristic	(N=52)		(N=25)		(N=27)				
	N	**%**	**N**	**%**	**N**	**%**	**Chi²**	**df**	***p***
Intervention condition							1.57	2	0.455
CBT	21	40.38	10	40.00	11	40.74			
ST	15	28.85	9	36.00	6	22.22			
IST	16	30.77	6	24.00	10	37.04			
Gender (female)	30	57.69	15	60.00	15	55.56	<0.00	1	0.966
Nationality (german)	44	84.62	22	88.00	22	81.48	1.07	2	0.586
Qualification for University entrance	27	51.92	14	56.00	13	48.15	2.96	2	0.227
Income							6.21	3	0.102
Low income	20	38.46	8	32.00	12	44.44			
Middle income	19	36.54	11	44.00	8	29.63			
High income	9	17.31	6	24.00	3	11.11			
not specified	4	7.69	0	0.00	0	0.00			
	M	***SD***	***M***	***SD***	***M***	***SD***	**t or W**	**df**	***p***
Age (years)	40.5	11.71	43.12	12.04	38.07	11.06	-1.57	48.70	0.123
EMA response rate	64.67	23.35	67.84	20.44	61.73	25.80	-0.95	48.87	0.347
Baseline Symptoms (BDI-II)	32.58	8.50	32.84	8.42	32.33	8.73	-0.21	49.91	0.832

Note. ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; BDI-II: Beck's Depression Inventory II; EMA: Ecological Momentary Assessment.

Research Questions 1a and 1b

To test whether early improvement predicts treatment response within the first four treatment weeks regardless of whether early improvement is assessed with EMA or WQA, we fitted logistic regression models, predicting treatment response by baseline depression and early improvement rates in depressive symptoms assessed via EMA versus WQA. Intervention condition was not a significant covariate in our models and therefore not included in further analyses. For each variable (EMA and WQA), we ran four different models denoting the four time windows (Table 3.2). For both EMA and WQA, early improvement rates of depressive symptoms significantly predicted treatment responder status not after one or two weeks of treatment (one week: EMA: z = -0.38, 95% CI = [-1.79 – 1.14], p = 0.704; WQA: z = -1.67, 95% CI = [-7.14 – 0.15], p = 0.095; two weeks: EMA: z = -0.84,

95% CI = [-1.59 - 0.52], p = 0.399; WQA: z = -1.87, 95% CI = [-5.83 - -0.15], p = 0.062), but after three and four weeks of treatment (three weeks: EMA: z = -2.33, 95% CI = [-4.49 - -0.51], p = 0.02; WQA: z = -2.32, 95% CI = [-7.65 - -0.94], p = 0.02; four weeks: EMA: z = -2.43, 95% CI = [-4.09 - -0.62], p = 0.015; WQA: z = -3.11, 95% CI = [-10.53 - -2.7], p = 0.002). These results indicate that early improvement in depressive symptoms, as assessed by both EMA and WQA, is a significant predictor of treatment response after three to four weeks of treatment. Figure B.2 in the Supplementary Material shows the improvement rate distributions of the four time windows separately for responders and non-responders and the two variables EMA and WQA.

Table 3.2

Logistic Regression Models Predicting Treatment Response based on Early Improvement Rates measured within four different Time Windows with EMA or WQA

DV/predictors	**Estimates**	**SE**	**z (*p*)**	**95% CI**	**AIC**	**BIC**
Early Improvement in EMA – Depressive symptoms - 1 week						
Intercept	-0.26	1.12	-0.23 (0.817)	[-2.52 - 1.95]	77.82	83.67
Baseline BDI	0.01	0.03	0.17 (0.866)	[-0.06 - 0.07]		
Early Improvement	-0.27	0.72	-0.38 (0.704)	[-1.79 - 1.14]		
Early Improvement in EMA – Depressive symptoms - 2 week						
Intercept	-0.16	1.13	-0.15 (0.884)	[-2.43 - 2.07]	77.19	83.05
Baseline BDI	0	0.03	0.09 (0.93)	[-0.06 - 0.07]		
Early Improvement	-0.43	0.51	-0.84 (0.399)	[-1.59 - 0.52]		
Early Improvement in EMA – Depressive symptoms - 3 week						
Intercept	-0.14	1.17	-0.12 (0.906)	[-2.5 - 2.19]	71.52	77.37
Baseline BDI	-0.01	0.04	-0.15 (0.883)	[-0.08 - 0.06]		
Early Improvement	-2.34	1	-2.33 (0.02)	[-4.49 - -0.51]		
Early Improvement in EMA – Depressive symptoms - 4 week						
Intercept	0.51	1.22	0.42 (0.678)	[-1.91 - 2.97]	69.49	75.34
Baseline BDI	-0.02	0.04	-0.62 (0.537)	[-0.1 - 0.05]		
Early Improvement	-2.12	0.87	-2.43 (0.015)	[-4.09 - -0.62]		
Early Improvement in WQA – BDI - 1 week						
Intercept	0.3	1.24	0.24 (0.807)	[-2.15 - 2.77]	74.49	80.35
Baseline BDI	-0.02	0.04	-0.52 (0.602)	[-0.1 - 0.05]		
Early Improvement	-3.05	1.83	-1.67 (0.095)	[-7.14 - 0.15]		
Early Improvement in WQA – BDI - 2 week						
Intercept	0.04	1.24	0.03 (0.977)	[-2.43 - 2.5]	73.6	79.46
Baseline BDI	-0.02	0.04	-0.46 (0.643)	[-0.09 - 0.06]		
Early Improvement	-2.67	1.43	-1.87 (0.062)	[-5.83 - -0.15]		
Early Improvement in WQA – BDI - 3 week						
Intercept	-0.61	1.22	-0.5 (0.615)	[-3.12 - 1.76]	70.89	76.75
Baseline BDI	-0.01	0.04	-0.37 (0.709)	[-0.09 - 0.06]		
Early Improvement	-3.94	1.7	-2.32 (0.02)	[-7.65 - -0.94]		
Early Improvement in WQA – BDI - 4 week						
Intercept	-1.92	1.37	-1.4 (0.161)	[-4.8 - 0.71]	62.93	68.78
Baseline BDI	0	0.04	-0.11 (0.911)	[-0.08 - 0.07]		
Early Improvement	-6.11	1.97	-3.11 (0.002)	[-10.53 - -2.7]		

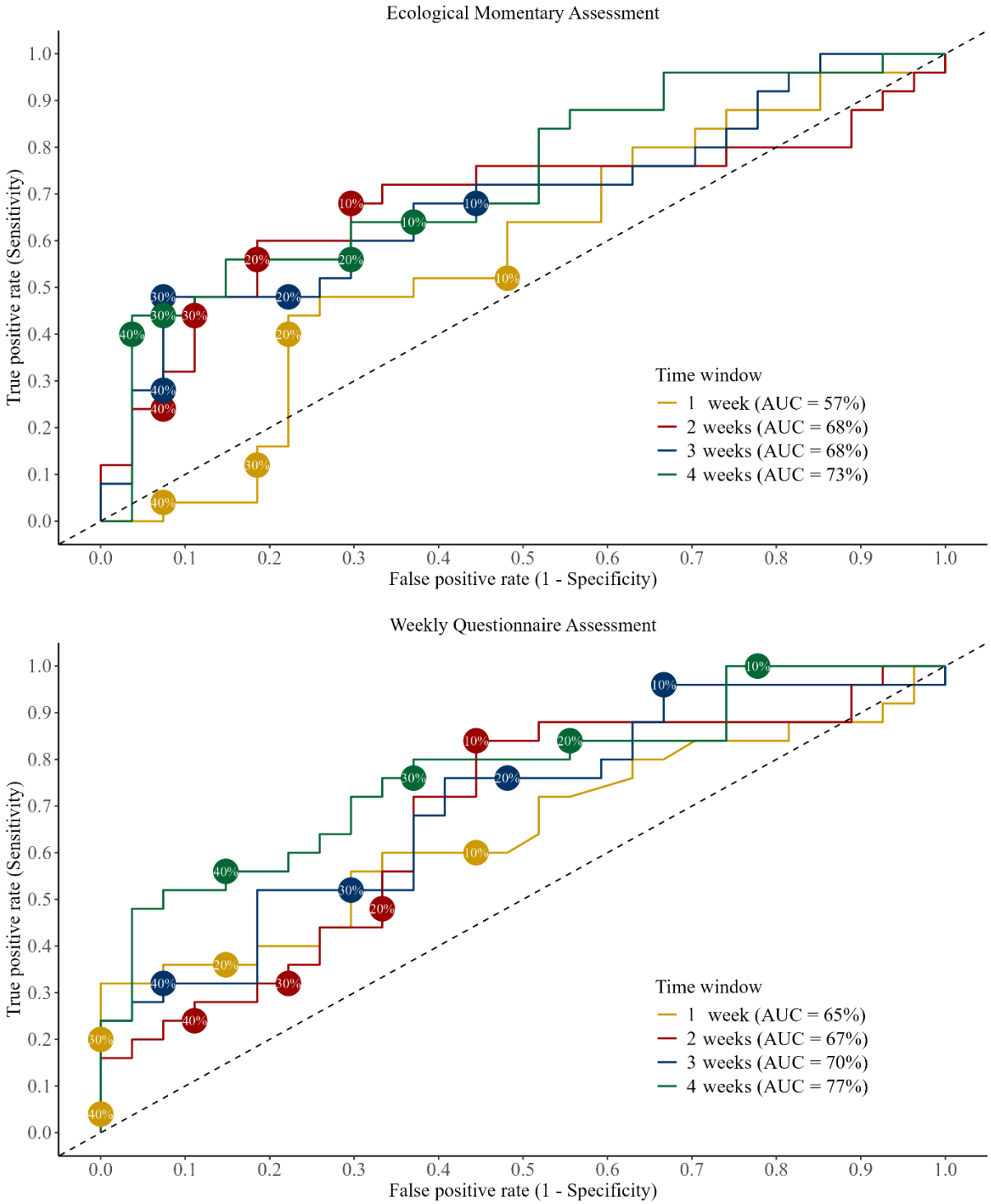
Note. EMA: Ecological Momentary Assessment; WQA: Weekly questionnaire assessment; Sample size in all models is 52. Early Improvement is defined as Improvement Rate.

Research Question 2

Next, we computed ROC-analysis and calculated the AUC values of each combination of variable (EMA versus WQA) and time window (one, two, three, or four weeks). All four time windows yielded AUC values exceeding 50%, indicating predictions better than chance on average (Figure 3.2). The highest AUC value was reached by WQA after four weeks of treatment (AUC = 0.77), which was slightly but not significantly larger than its AUC value after three weeks of treatment (AUC = 0.7; p = 0.143, 95% CI = [-0.17 – 0.02]) or the AUC value of EMA after four weeks of treatment (EMA: AUC = 0.73; *p* = 0.59, 95% CI = [-0.19 – 0.11]). In summary, these results indicate that both WQA and EMA provide comparable predictions, which are better than chance and become particularly significant after three to four weeks of treatment. For the sake of completeness and comparability, comparison analyses between all AUC values and the maximum youden-index of each time-window are reported in Table B.4 in the Supplementary Material.

Figure 3.2

Receiver Operating Characteristic (ROC) Plots for the Prediction of Treatment Response after 7 weeks through Early Improvement in Depressive Symptoms measured by Ecological Momentary Assessment or Weekly Questionnaire Assessment within four different Time windows



Note. *n* =52. The ROC curves illustrate the prediction accuracy of early improvement measured with EMA and WQA within four distinct time windows: first treatment week =

yellow, first two weeks = red, first three weeks = blue, and first four weeks = green. Dots mark specifically investigated rates of early improvement (≥ 10%, 20%, 30%, and 40%).

Finally, we explored which combination of time window and symptom cutoff yields the 'best' definition of early improvement. Figure 3.3 shows the distribution of true positive, true negative, false negative and false positive predictions for each definition. Additionally, for each definition we calculated (Table 3.3): (a) the sensitivity, (b) the specificity, (c) the youden-index (y), i.e., the percentage of correct predictions, (d) our predefinied TNFN ratio, i.e., the ratio between the true negative to false negative rate (specificity/(1-sensitivity)), and (e) the weighted TNFN ratios (wTNFN), taking the responder rate found in our sample of RR = 48% into account (wTNFN ratio (RR) = TNFN ratio*((1-RR)/RR)). All but two definitions of early improvement yielded predictions better than chance (y > 0) and wTNFN ratios >1, indicating higher numbers of true negative to false negative predictions. The 'best' definition of early improvement was a WQA-assessed 10% improvement by week four, reaching a true negative rate (specificity) of 22% in combination to a false negative rate of 0%. Accordingly, the TNFN ratio and wTNFN ratio resulted in infinite values. Specifically, this definition of early improvement resulted in 48% true positive (n = 25), 12% true negative (n = 6), 0% false negative (n = 0) and 40% false positive (n = 21) predictions in our sample (Table 3.3).

Table 3.3

Prediction Metrics of different Definitions of Early Improvement in Depressive Symptoms measured by EMA or WQA on Treatment Response to a seven-week Psychological Treatment

	(Specificity, Sensitivity, Youden-Index), TNFN, wTNFN,			
	10% improvement	**20% improvement**	**30% improvement**	**40% improvement**
Early Improvement assessed with Ecological Momentary Assessment				
1 week	(52%, 52%, 0.04) TNFN = **1.08** wTNFN = 1.17	(40%, 78%, 0.18) TNFN = **1.3** wTNFN = 1.4	(12%, 81%, -0.07) TNFN = **0.93** wTNFN = 1	(4%, 93%, -0.03) TNFN = **0.96** wTNFN = 1.04
2 weeks	(68%, 70%, 0.38) TNFN = **2.2** wTNFN = 2.38	(56%, 81%, 0.37) TNFN = **1.85** wTNFN = 2	(44%, 89%, 0.33) TNFN = **1.59** wTNFN = 1.71	(24%, 93%, 0.17) TNFN = **1.22** wTNFN = 1.32
3 weeks*	(68%, 56%, 0.24) TNFN = **1.74** wTNFN = 1.88	(48%, 78%, 0.26) TNFN = **1.5** wTNFN = 1.62	(48%, 93%, 0.41) TNFN = **1.78** wTNFN = 1.92	(28%, 93%, 0.21) TNFN = **1.29** wTNFN = 1.39
4 weeks*	(64%, 63%, 0.27) TNFN = **1.75** wTNFN = 1.89	(56%, 70%, 0.26) TNFN = **1.6** wTNFN = 1.73	(44%, 93%, 0.37) TNFN = **1.65** wTNFN = 1.79	(40%, 96%, 0.36) TNFN = **1.6** wTNFN = 1.73
Early Improvement assessed with Weekly Questionnaire Assessment				
1 week	(60%, 56%, 0.16) TNFN = **1.39** wTNFN = 1.5	(36%, 85%, 0.21) TNFN = **1.33** wTNFN = 1.44	(20%, 100%, 0.2) TNFN = **1.25** wTNFN = 1.35	(4%, 100%, 0.04) TNFN = **1.04** wTNFN = 1.13
2 weeks	(84%, 56%, 0.4) TNFN = **3.47** wTNFN = 3.75	(48%, 67%, 0.15) TNFN = **1.28** wTNFN = 1.38	(32%, 78%, 0.1) TNFN = **1.14** wTNFN = 1.24	(24%, 89%, 0.13) TNFN = **1.17** wTNFN = 1.26
3 weeks*	(96%, 33%, 0.29) TNFN = **8.33** wTNFN = 9	(76%, 52%, 0.28) TNFN = **2.16** wTNFN = 2.33	(52%, 70%, 0.22) TNFN = **1.47** wTNFN = 1.58	(32%, 93%, 0.25) TNFN = **1.36** wTNFN = 1.47
4 weeks*	(100%, 22%, 0.22) TNFN = ∞ wTNFN = ∞	(84%, 44%, 0.28) TNFN = **2.78** wTNFN = 3	(76%, 63%, 0.39) TNFN = **2.62** wTNFN = 2.83	(56%, 85%, 0.41) TNFN = **1.94** wTNFN = 2.09

Note. *n* = 52; The table compares the following definions of early improvement in predicting treatment response: ≥ 10%, 20%, 30%, or 40% improvement after one, two, three, or four weeks of treatment, measured with Ecological Momentary Assessment or Weekly Questionnaire Assessment. In brackets the metrics sensitivity, specificity, and youden-index are reported. TNFN denotes the ratio of the true negative to false negative rate (specificity/1-sensitivity). wTNFN denotes the weighted TNFN ratio taking the responder rate (RR) of the investigated sample, which was 48% into account (TNFN*((1-RR)/RR)). Asterisks (*) highlight time windows that significantly predicted treatment response (for details, see Table 3.2). The cells colors indicate the size of the TNFN ratios (bold) in relation to each other reaching from light green (lower values) to dark green (higher values).

Figure 3.3

Rates of False and Correct Predictions of Treatment Response based on different Definitions of Early Improvement

Note. *n* = 52. Following definitions of early improvement were examined: an improvement in depressive symptoms of ≥ 10%, 20%, 30% or 40% after one, two, three or four weeks of treatment as measured by EMA or WQA. All but one definition predict treatment response better than chance (dashed line). For clinical decision-making it is particularly important that

the number of true negatives (light green) is high and, above all, higher than the number of false negatives (light red).

Discussion

The aim of this study was to investigate whether early improvements in depressive symptoms measured by both EMA and WQA could significantly predict treatment response versus non-response and which specific definitions of early improvement may be most useful to support clinical decisions.

Treatment Prediction through EMA and WQA

Confirming our first hypotheses, we found that early improvements assessed by both EMA and WQA serve as robust predictors of treatment response. Three weeks after treatment initiation, both assessment techniques reached significant predictions of responder status. This corresponds to a treatment dose of 12 sessions (two individual and two group sessions per week) and about halftime in our study setup. We observed that the prediction accuracies of EMA and WQA did not significantly differ from each other. This underlines the predictive power of EMA. This is especially notable as the early improvement assessed via WQA and the outcome assessment were homogenous in measurement (both used the BDI-II), whereas the EMA assessment differed from the outcome assessment in both general methodology (EMA vs. retrospective assessment) as well as specific items used.

Our results align with previous studies showing early improvement after four weeks of treatment to be predictive (Beard & Delgadillo, 2019). However, given the intensity and type of the comprehensive psychiatric care program, including two individual and two group sessions of psychotherapy per week, it would have also been reasonable to expect significant predictions as early as one or two weeks into treatment. Studies on early change patterns show that most change in patients' progress occurs early in treatment e.g., within the first four sessions (Rubel et al., 2015), which corresponds to a treatment duration of one week in our study setup. Besides, studies on dose-effect relationships show that frequent treatment schedules, like in our study (two group and two single sessions) are more effective for the treatment of depression than schedules with only one session per week (Cuijpers et al., 2013). However, although early improvement within the initial weeks of treatment is common, this does not necessarily mean that it is also predictive for the final treatment outcome, as it could be influenced by non-specific factors related to study participation, such as expectation biases (Targum et al., 2020). Besides, Lutz et al. (2017) showed that the pattern of patients' early

improvement adds beneficial information to the simple measure of 'how much' they improve. Specifically, they found three different patterns of early changes within the initial six weeks of a web-based intervention study for the treatment of depression. The trial started with a screening at baseline and a registration after 2 weeks, followed by 4 weeks of intervention. The three patterns found were: continuous improvement starting directly after screening, improvement starting after registration, i.e., with a two-week delay, and linear deterioration starting directly after screening. Interestingly, the delayed improvement was the most accurate predictor of treatment outcome, followed by the direct improvement. Therefore, the methodological approach of our study, in which we estimated change rates using linear regressions involving multiple assessment points rather than simply calculating pre-post differences between two time points, may have positively impacted the accuracy of our predictions of both EMA and WQA.

Early Improvement Definition

The second aim of our study was to explore different definitions of early improvement, combining the four different time windows with four different symptom cutoffs. Comparing the AUC values of our ROC-analyses revealed that the significant prediction accuracy of the time window three weeks, did not significantly improve when it was extended to four weeks. Additionally, the ROC-analyses revealed that almost all investigated definitions of early improvement yielded prediction accuracies above chance, as indicated by positive youden-indices (Table 3.3). For the symptom cutoffs, we observed that WQA reached the highest ratios between the true negative and false negative predictions with a symptom improvement of ≥ 10%. However, both EMA and WQA achieved ratios > 1, indicating more true negative than false negative predictions.

This suggests that small cutoffs such as a ≥ 10% improvement assessed with WQA three to four weeks after treatment initiation might serve as a good predictor of treatment response that could guide clinicians in their decision about whether to continue or change a psychological treatment in the sense of stepped-care or modularized therapy. The definition identified responders with a sensitivity of 100% and non-responders with a specificity of 22% without taking any false negative prediction into account. This means after four weeks in our study, 0 patients (0%) would have been falsely recommended to change their treatment, while 6 patients (12%) would have been correctly advised to do so and therefore spent 3 weeks (42%) less in an ineffective treatment before considering alternatives. In total, this would

have saved 18 weeks (~ 4.5 months) that our clinicians spend treating patients without the prospect of a response. From a clinical and economic perspective, looking at the ratio between the true negative and false negative predictions is a promising approach to push the development of decision rules for treatment allocation in the sense of stepped-care and personalized therapy.

Limitations

Despite the robustness of our findings, several limitations warrant consideration. Firstly, the lack of validation for the identified definitions of early improvement underscores the need for further research to confirm our results. Additionally, the modest sample size and unique therapy setting of our study, involving intensive inpatient or day-clinic treatment with four therapy sessions per week and optional concomitant care, may limit the generalizability of our findings to other treatment contexts. Future studies should aim to replicate our findings across diverse treatment durations and intensities, less severe levels of depression, and different settings. Moreover, the results need to be validated in a sample without concomitant care.

Exploring alternative methods for assessing responder status could provide deeper insights into the comparative effectiveness of EMA and WQA assessments. In our study, we measured treatment response with the same measure (namely the BDI-II) as the WQA-assessed early improvement rates. This confounds the comparisons between our EMA and WQA predictions, as measurement instruments may predict themselves better than others. Moreover, this opens the question which measurement instrument might generate the most valid and reliable assessment of change in depressive symptoms. When comparing EMA and WQA, it is also important to consider that multiple EMA prompts per day might cause higher patient burden than WQA (van Genugten et al., 2020). Therefore, it could be promising to develop EMA approaches that track depression-related behaviors passively by using wearables (e.g., sleep, activity, stress) and formulate EMA items neutrally to avoid systematically drawing patients' attention to negative aspects. Moreover, the substantial proportion of missingness in our sample is a limitation of our study. More than 50% of the originally recruited sample had to be excluded due to dropout and/or missing data, which limits our conclusions to patients who completed our treatment and assessments.

Finally, it is important to note that our wTNFN ratio depends on the responder rate of the population. As our prediction analyses were retrospective, we knew the exact responder

rate of our sample. In a prospective study, the responder rate would have to be estimated. Especially when high responder rates are expected, it is important to prove which TNFN ratios reach their target. E.g., when our TNFN ratios are transferred to a sample with 90% responders, the TNFN ratio of a WQA-assessed 10% improvement after week three of TNFN = 8.33 results in a wTNFN <1 (wTNFN(RR=0.9): 8.33*((1-0.9)/0.9) = 0.93). This means that in this sample the TNFN ratio would not have resulted in more true negative than false negative predictions as the rate of false negative predictions increases proportionally to the rate of responders.

Conclusion

In conclusion, our study highlights the potential of early improvements in depressive symptoms, measured through EMA or WQA, serving as robust predictors of treatment response. The insight that after three weeks of treatment WQA was able to detect non-responders with a specificity of 33% while taking a false negative rate of only 4% into account, coupled with the importance of small cutoffs to yield such high ratios between the specificity and false negative rate, offers valuable implications for further research and the development of clinical decision-rules. Moreover, they can be used in trial designs that aim to speed up the development of psychological treatments such as the ‘leapfrog’ method (Blackwell et al., 2019). By identifying early predictors of treatment response, our findings could help streamline this process, making it more efficient. The validation of EMA as an alternative tool to WQA inspires further research investigating its potential when different definitions of treatment response are used. However, we recognized that there is little research on how effective non-responders of one intervention can be treated with another, which is an important assumption of treatment prediction research in any way. Finally, further validation studies are warranted to confirm our findings and support their integration into clinical practice.

4. Study III:

The Role of Concreteness in Repetitive Negative Thinking: Temporal Dynamics and the Predictive Value for Depression Throughout Psychological Treatment

This chapter is a pre-peer-review, pre-copyedit version of a manuscript submitted to *Behaviour Research and Therapy*.

Kirchler, S. V., Müller, C. L., Spock, Z., Ehring, T., Kopf-Beck, J.* & Tamm, J.* (under review). The Role of Concreteness in Repetitive Negative Thinking: Temporal Dynamics and the Predictive Value for Depression Throughout Psychological Treatment. Manuscript submitted to *Behaviour Research and Therapy.*

** The last two authors contributed equally to this work.*

Abstract

Introduction: Repetitive negative thinking (RNT) is an important transdiagnostic process involved in the development and maintenance of depression. Evidence suggests that maladaptive RNT is characterised by reduced concreteness. However, the temporal relationship between concreteness of RNT and depressive symptoms, as well as changes in concreteness during psychological treatment, remain unclear. Therefore, the current study investigated (a) whether momentary RNT concreteness explains variance in the prediction of momentary depressive symptoms beyond momentary RNT, (b) whether momentary RNT concreteness increases over the course of psychotherapy and (c) the temporal precedence between momentary RNT concreteness and momentary depressive symptoms. **Methods**: Seventy-seven depressed patients participating in a randomised controlled trial were assessed using Ecological Momentary Assessment (EMA) during a seven-week inpatient psychological treatment. EMA, conducted three times daily, included measures of depression, RNT, and a free-text item assessing patients' RNT thoughts, which were rated for concreteness by trained raters. Weekly depression severity was assessed using the Beck Depression Inventory-II. Hypotheses were tested using multilevel modelling. **Results**: Concreteness of RNT was significantly associated with depression. A model incorporating both RNT and concreteness accounted for significantly more variance in depression than a model with RNT alone. Concreteness of RNT increased throughout treatment, dependent on patients' improvement in depression severity. Depression levels predicted subsequent concreteness, but not vice versa. **Discussion**: Concrete thinking is consistently related to depression and improves over the course of effective psychological treatment. However, the current findings do not suggest that changes in concreteness predict subsequent reduction of depression. levels. Future research should explore long-term temporal dynamics between RNT concreteness and depression to evaluate the potential of concreteness as a mechanism of change in psychological treatments in more detail.

Introduction

Repetitive negative thinking (RNT) is a transdiagnostic process involved in the development and maintenance of depression and other emotional disorders (Watkins et al., 2012). RNT is thereby defined as a cognitive process of recurrent thinking about negative content that is typically experienced as intrusive and difficult to control (Ehring & Watkins, 2008). It includes the subprocesses of worry and rumination, with the latter being specifically linked do depression (Watkins & Roberts, 2020). Worry, defined as “a chain of thoughts and images, negatively affect-laden and relatively uncontrollable” (Borkovec et al., 1983, p. 9) focuses on a potential negative event in the future (Borkovec et al., 1991) while depressive rumination (Nolen-Hoeksema, 1991), described as “repetitive thinking about the symptoms, causes, circumstances, meanings, implications and consequences of depressed mood” (Watkins & Roberts, 2020, p. 1). focuses on past events (Watkins et al., 2005).

Several mechanisms have been suggested to be involved in the development of maladaptive RNT, including its negative valence (Segerstrom et al., 2003) and its habitual nature (for an overview see Watkins & Roberts, 2020). The concreteness theory of worry (Stöber, 1998) furthermore postulates that the maladaptive character of worry is highly associated with its lack of concreteness. This was confirmed by a large body of studies (e.g. Altan-Atalay et al., 2022; Behar et al., 2012; Stöber, 2000; Stöber & Borkovec, 2002; Watkins, 2008; Watkins & Roberts, 2020).

Abstract worrying, defined as "indistinct, cross-situational, equivocal, unclear and aggregate" (Stöber & Borkovec, 2002, p. 92), interferes with problem solving (Stöber, 1998), functions as a maladaptive coping strategy by inhibiting the integration of anxiety-incongruent information (Foa & Kozak, 1986), and ultimately facilitates and maintains symptoms of emotional disorders. On the other hand, concrete thinking, described as "distinct, situationally specific, unequivocal, clear and singular" (Stöber & Borkovec, 2002, p. 92) serves as an adaptive coping strategy that reduces RNT and associated psychopathology (Stöber & Borkovec, 2002).

Building upon this theory, the concreteness theory of worry has also been increasingly researched in relation to the rumination aspect of RNT. Reduced concreteness has been shown to be particularly associated with higher levels of depression (Kircanski et al., 2015, Watkins & Moulds, 2005, 2007; Takano & Tanno, 2010). Moreover, concreteness has been shown to play an important role in the interplay between rumination and depressive symptoms. Takano and Tanno (2010) observed that momentary ruminative self-focus is associated with

concurrent negative affect only when concreteness is low. Furthermore, concrete versus abstract emotion differentiation has been posited to be associated with concreteness in thinking (Liu, Gilbert & Thompson, 2020). Specifically, it has been shown to moderate the association between rumination and depressive symptoms (Starr et al., 2017) through the ability to recognize and concretely describe emotions.

Although there is strong evidence for an important role of rumination and concreteness in depression, the question of the temporal relationship between them remains open. Nolen-Hoeksema (1991) proposed in her Response Styles Theory that rumination is a response to negative mood, which then functions as a maintenance factor leading to longer periods of depressive symptoms (Nolen-Hoeksema, Morrow, & Fredrickson, 1993) and as a risk factor predicting new-onset depression (Nolen-Hoeksema, 2000). Similarly, rumination at a previous timepoint was shown to predict an increase in negative affect and a decrease in positive affect (Kircanski, Thompson, Sorenson, Sherdell, & Gotlib, 2018), and negative affect, in turn, increased rumination at the subsequent timepoint (Moberly & Watkins, 2008). Conversely, concrete (but not abstract) positive memories are followed by mood repair in depressed and recovered depressed individuals (Werner-Seidler & Moulds, 2012).

The important role of concreteness and rumination in depression makes it an ideal target for intervention. Concreteness training has been shown to be effective in increasing concrete thinking and decreasing depressive symptoms (e.g., Watkins & Moberly, 2009; Watkins et al., 2012). It includes psychoeducation about RNT, experiential exercises, and strategies for shifting from abstract to more concrete thinking with tools for implementation in everyday life. It can be delivered, for example, as a guided self-help intervention in addition to usual care to reduce depressive symptoms (Watkins et al., 2012), as a mobile app for the prevention of depression and anxiety disorders in adolescents (Funk et al., 2023; Funk et al, 2024; Funk et al., 2025) and for reducing current RNT and symptoms of anxiety and depression (Edge et al., 2024), or as part of rumination-focused cognitive behavioural therapy (rfCBT; Watkins, 2016; Wallsten et al., 2023). Although common psychological interventions for depression often do not explicitly include RNT-focused strategies, it is conceivable that they may also have an impact on concrete thinking. For example, cognitive interventions guide patients to challenge their negative thoughts and modify them in a very concrete way. Similarly, cognitive behavioural (Beck, 2021) or problem solving-focused interventions (Nezu et al., 2013) ask patients to very concretely think about problems, make and follow plans, and evaluate their success. In line with these considerations, Stöber and Borkovec

(2002) found initial evidence that the concreteness of worry increased significantly after cognitive-behavioural therapy (CBT) in patients with generalized anxiety disorder. These findings raise the question of whether common psychological interventions have similar effects in depressed patients and how the temporal dynamics between concreteness and depressive symptoms evolve throughout treatment.

The present study

To better understand the role of concrete thinking and the temporal dynamics of changes in concreteness, RNT, and depressive symptoms, it is necessary to go beyond pre-post measures and to use high-frequency measures such as Ecological Momentary Assessment (EMA). In order to delineate these dynamics over the course of different psychological interventions, it is mandatory to apply EMA throughout the whole treatment period. Therefore, we used EMA in a clinical sample of severely depressed patients during a seven-week inpatient treatment to answer three main research questions: First, whether concreteness adds explanatory power to the prediction of depressive symptoms (beyond RNT) in depressed individuals; second, whether non-specific psychiatric care (including psychological treatments) is effective in improving concreteness levels of thinking; and third, how the temporal dynamics between concreteness of RNT on the one hand and depressive symptoms on the other unfold.

The present study is part of the OPTIMA trial, a mono-centric, parallel-group, block-randomized controlled trial (RCT) designed to evaluate the effectiveness of schema therapy for depression (Egli et al., 2019). The trial employs a superiority design, comparing schema therapy to a non-specific individual supportive therapy, which has been utilized in previous psychotherapy trials (Schramm et al., 2011). Additionally, schema therapy is compared to cognitive behavioural therapy (CBT, Hautzinger, 2013). A particular focus of the study is the investigation of mechanisms of change across all treatment conditions. All psychotherapeutic interventions were integrated into a comprehensive inpatient or day-clinical treatment framework. Each treatment followed a manual and all therapist received training and regular supervision. Diagnosis of depression and comorbidities were assessed using the Munich-Composite International Diagnostic Interview (M-CIDI, Wittchen et al., 1998). All clinical ratings were conducted by blinded raters. To clarify the role of RNT and its temporal dynamics with depressive symptoms, continuous EMA was implemented. For details on design, measures and objectives of the trial and the current research see the study protocol (Kopf-Beck et al., 2020).

To our best knowledge, the current study is the first to investigate changes of RNT concreteness assessed continuously by trained raters over the course of treatment, while previous studies have primarily focused on general differences between levels of RNT concreteness of depressed, recovered and healthy individuals (Watkins & Moulds, 2007), depression levels within the appliance of concreteness trainings (Watkins & Moberly, 2009) and rfCBT (Wallsten et al., 2023) or without treatment (Takano & Tanno, 2010).

We hypothesised that (a) momentary RNT concreteness explains variance in the prediction of momentary depressive symptoms beyond momentary RNT, and (b) momentary RNT concreteness increases over the course of psychotherapy. In addition, we further explored (c) the temporal relationship between momentary RNT concreteness and momentary depressive symptoms.

Materials and Methods

Design and Procedures

The current study was conducted with a subsample of patients participating in the OPTIMA study (Kopf-Beck et al., 2020), run at the Max Planck Institute of Psychiatry in Munich, Germany. The study was approved by the Institutional Ethic Committee of the Faculty of Medicine at LMU Munich (Project number 17-395). The study was conducted in accordance with the Declaration of Helsinky and all patients provided informed consent prior to inclusion. Patients received a reimbursement based on their adherence rate on the EMA.

Participants

Inclusion criteria of the OPTIMA study were (1) age between 18 and 75 years, (2) a diagnosis of major depressive disorder (single or recurrent episode, moderate or severe) represented by ICD-10 diagnoses (F32.1, F32.2, F33.1, or F33.2), diagnosed by clinical assessment and indicated by a score greater than or equal to 20 on the baseline survey of self-reported depression by the Beck Depression Inventory II (BDI-II, Hautzinger et al., 2006) or the Montgomery-Åsberg Depression Rating Scale (MADRS, Montgomery & Asberg, 1979). To take part in the EMA sub study, patients additionally (3) had to be in possession of a smartphone that was compatible with the EMA application.

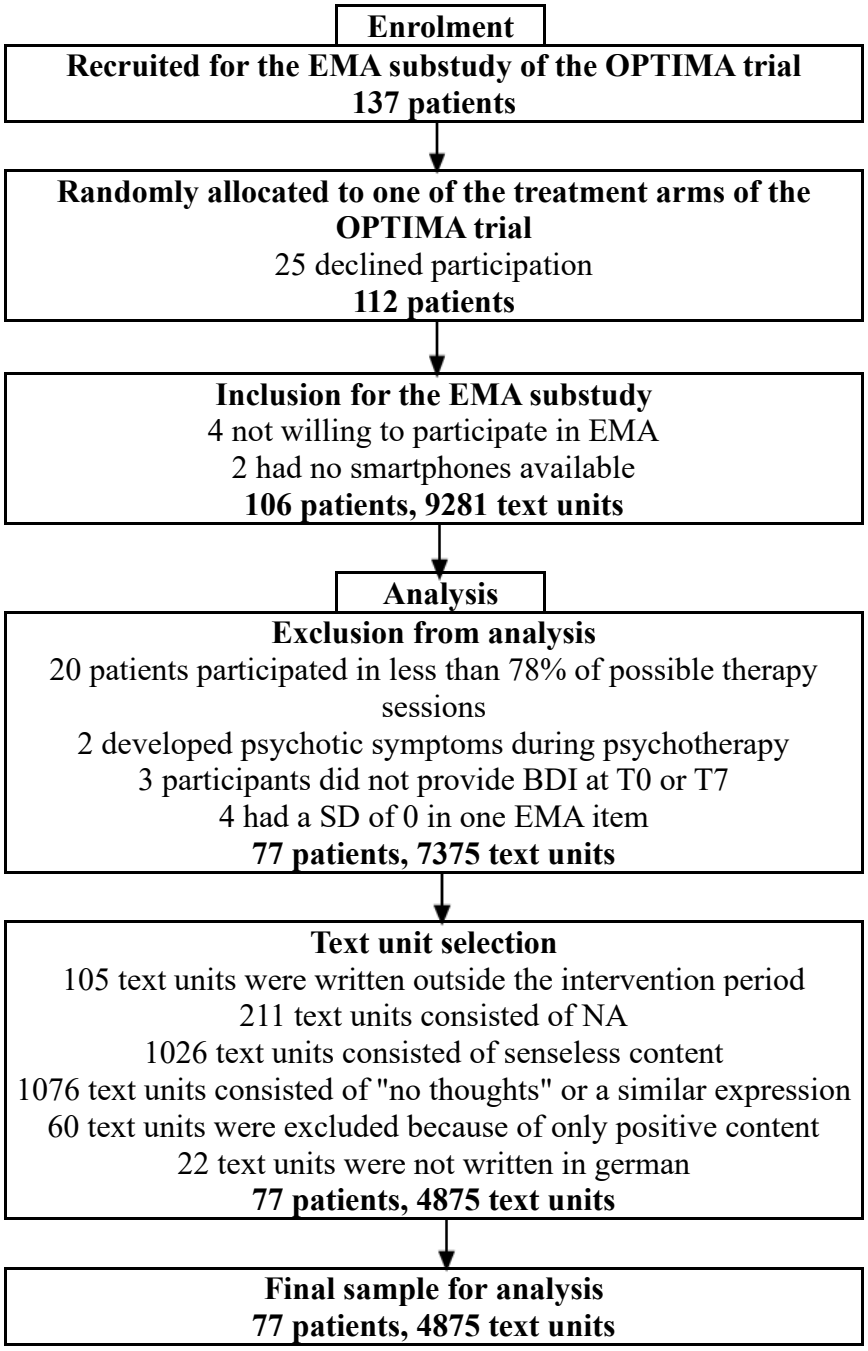
The exclusion criteria used in our study were also adopted directly from the larger RCT (OPTIMA trial), in which our EMA substudy was embedded. Patients were excluded in the presence of (1) acute suicidality, (2) lifetime history of psychotic or bipolar disorder, (3)

severe concomitant neurological or internal diseases, (4) organic mental disorders, (5) acute alcohol abuse, (6) withdrawal symptoms from illicit drugs, (7) substance-induced disorder, (8) severe mutism or stupor, (9) mental disorder secondary to a concomitant medical condition or secondary to a substance use disorder, (10) the presence of a language barrier that prevented participation in psychotherapy, (11) an IQ under 80 or a severe learning disability. Additionally, (12) women during pregnancy or lactation were excluded aligning with established clinical trial standards aimed at minimizing risks to the child and safeguarding maternal health.

After checking on the OPTIMA exclusion criteria, participants were excluded from further analyses (1) if data on depression severity (BDI-II) were missing at pre- or post-treatment or (2) if they had a standard deviation of zero in at least one EMA item over the whole course of treatment. Excluding patients with no variance in their responses to at least one EMA item over a long assessment period (in our case, seven weeks) is a data quality approach that has been used in previous EMA studies (e.g., Rosenkranz et al., 2020; Tamm et al., 2024) to identify patients who may not have conscientiously responded to the EMA assessment. Consistent with the study protocol (Kopf-Beck et al., 2020), (3) drop-outs (patients who participated in less than 78% of the therapy sessions or who were identified as meeting exclusion criteria during the study such as the development of psychotic symptoms) were also excluded from the data analysis. Moreover, further quality assurance was implemented in the data cleaning. This was required particularly for one of the EMA-items, which consisted of a free-text format and whose free-text responses are referred to as “text units” in the following. We excluded text units that (4) were written outside the intervention period, (5) empty entries (NA), and (6) that contained no letters, or interpretable content (e.g., "4§$][" or “Walking potato yesterday”). Further, we excluded text units (7) expressing no RNT (e.g. "no thoughts") or (8) not answering the question about RNT of the EMA-item (e.g. referring only to positively valanced content). Moreover, text units that were (9) not written in German were excluded. The remaining sample consisted of 77 participants with 4875 text units. An overview of the number of excluded participants and text units can be found in Figure 4.1.

Figure 4.1

Data exclusion flow diagram.



Note. The exact exclusions of the cleaning process are displayed in the above flow diagram.

Regarding the remaining sample, a mean of *M*=11.98 (*SD*=15.23) words per text unit was written and a mean of *M*=63.31 (*SD*=48.19) text units per person was assessed. A detailed overview of text unit frequency can be found in Figure C.1 in the Supplementary Material.

Interventions

Participants were randomly allocated to one of three psychological intervention conditions: CBT, ST or IST. After baseline measurement (T0), patients participated in one of the three therapy conditions for up to seven weeks (T1-T7), including two individual treatment sessions per week (50 minutes per session), two group treatment sessions per week (100 minutes per session), and optional supportive psychopharmacological medication (e.g., antidepressants). In addition, patients could attend various additional interventions that are common in psychiatric inpatient or day clinic settings, such as occupational therapy, sports therapy or mindfulness exercises. For further details about randomisation and interventions see the OPTIMA study protocol (Kopf-Beck et al., 2020).

Measures

A detailed description of all measures conducted in the OPTIMA study can be found elsewhere (Kopf-Beck et al., 2020). In the following, only measures that are relevant to the conducted analysis are described. An overview of the assessment plan can be found in Table C.1 in the Supplementary Material.

Ecological Momentary Assessment

Throughout the entire intervention period, participants responded to three EMA prompts per day (morning, midday and evening). The time windows of the EMA prompts depended on the approximate wake-up times participants had indicated during the app onboarding process (morning = two hours before to five hours after wake-up time, midday = five to ten hours after wake-up time, and evening = ten hours after to two hours before wake-up time). Within each time window, participants received one semi-randomised reminder to complete the EMA prompt. Overall, n=168 prompts were provided including baseline (one baseline week + seven intervention weeks = 56 days*3 prompts per day). Within each EMA prompt, participants completed one open-text item capturing current RNT thoughts, which was subsequently rated for concreteness by trained human raters, as well as four items assessing momentary RNT, and four items measuring momentary depressive symptoms. The aggregation of the items into the three scores – RNT concreteness, momentary RNT, and momentary depressive symptoms – was conducted as follows:

Concreteness of RNT. The open-text item consisted of the following question (englisch translation): "Which negative thoughts are currently going through your mind repeatedly?

Please write down your thoughts in complete sentences". Participants could provide an answer in a free text field. After data collection, these text units were rated from two trained human raters according to Stöbers concreteness scale with the five categories abstract (1), somewhat abstract (2), neither-nor (3), somewhat concrete (4), and concrete (5) (Stöber & Borkovec, 2000). The use of external ratings is an established method in concreteness research (e.g., Stöber, 2000; Stöber & Borkovec, 2002; Watkins & Moulds, 2007). As outlined by Wahl et al. (2019), self-ratings of RNT concreteness may reflect patients' perceived concreteness, rather than objective ratings of their thoughts, overestimating concreteness levels and resulting in low variance. Since the text units were very heterogenic in length (ranging from 1 to 165 word per text unit), language and content, the authors SK and JT refined Stöber's (2002) concreteness scale to a decision aid to ensure a consistent interpretation of the scale. This contained checking for certain criteria, e.g. whether very generalising words such as "everything", "nobody", "always" etc. were mentioned, whether the situation was interpreted, whether solutions for problems were mentioned etc. (for details see: Figure C.2 in the Supplementary Material).

The rating procedure of the RNT concreteness score included a development, training and rating phase. First, the authors SK and JT formulated and tested the decision aid (Figure C.2 in the Supplementary Material) based on Stöber's concreteness scale in six feedback loops until a sufficient inter-rater reliability (Krippendorff's alpha) of > 0.7 for 150 rated text units was reached. Krippendorff (2019) suggests a minimum reliability of 0.667, with an alpha of 0.8 indicating good reliability. The decision aid follows Stöber's five-point Likert scale, ranging from 1 (abstract) to 5 (concrete). In the following training phase, two psychologists were trained as independent raters (CM and ZS) and rated randomly selected 1000 text units from the entire sample. As the two raters reached an inter-rater-reliability of > 0.7, the procedure was continued. Finally, the two raters rated all text units. Thereby, both raters were instructed to rate a maximum of n=1000 text units per week and to take regular breaks within each rating session to prevent exhaustion. All text units were pseudo-randomised by timepoint and participants prior to the ratings to prevent rating biases due to implicit assumptions about changes in concreteness over time or associations to depression severity. In this way, both raters were blind to the timepoints of the text units as well as characteristics of the writers, such as their depression severity and treatment condition. The final concreteness ratings of the two raters yielded an inter-rater reliability of 0.74 (Krippendorff, 2019). The final concreteness-score was calculated from the mean of the two raters' ratings per text unit.

Momentary RNT. To assess momentary RNT, we used an EMA paradigm developed and validated by Rosenkranz et al. (2020), which demonstrated excellent model fit, high reliability, and strong validity in predicting depression outcomes. The paradigm comprises four EMA items. Three items capture the core characteristics of RNT from the Perseverative Thinking Questionnaire (PTQ; Ehring et al., 2011): repetitiveness ("The same negative thoughts keep going through my mind again and again."), intrusiveness ("Negative thoughts come to my mind without me wanting them to."), and uncontrollability ("I get stuck on certain negative issues and can't move on."). The fourth item assesses the subjective burden through RNT ("I feel weighted down by negative thoughts.").

Momentary Depressive Symptoms. The momentary depressive symptoms score consisted of four items based on ICD-10 criteria for Major Depression, focusing on three core symptoms: lowering of mood (Item "Current Mood", "How are you feeling?"), reduction of energy and interest (Item "loss of interest", "Do you feel like you don't want to do anything anymore?"), and decrease in activity (Item "Withdrawal", "Are you currently withdrawing from social contacts or activities?"). Additionally, an item assessing psychomotor agitation or inhibition (Item "Psychomotor agitation/inhibition", "Are you feeling particularly physically inhibited or agitated?") was included to address somatic syndromes (ICD-10, F32, fifth position), which are particularly relevant in severe depression, especially in inpatient settings. The EMA-items were developed by the authors JKB and JT. The original German wordings of the momentary RNT and momentary depressive symptoms EMA items can be found in Table C.2 in the Supplementary Material.

To reduce patient burden, the response formats of the momentary RNT and momentary depressive symptoms items were two-stepped: First, participants determined whether an item was currently true or false (*yes/no*). Then, if true, the severity was rated on a five-point Likert scale (*not at all – very much*). We chose to present the full range of the five-point Likert scale in the second step, even though the option 'not at all' after an agreement in the first step is usually redundant. The item "current mood" constituted an exception and was evaluated using five emojis (labelling: *very good – good – moderate – bad – very bad*). The scores of momentary RNT and momentary depressive symptoms were constituted as the means of their respective four items, resulting in a possible range from 0 to 4. The reliability of the EMA scores for momentary RNT and momentary depressive symptoms showed good within-participant reliability (depressive symptoms: Rc=0.79, RNT: Rc=0.86) and excellent between-participant reliability (depressive symptoms: RkF=0.91, RNT: RkF=0.95; for details

see: Tamm et al., 2024). The study design of the EMA sub study was registered before the completion of data collection and prior to any analyses (osf.io/9fuhn).

BDI-II

To assess participants' change in depression severity throughout the trial, the German version of the revised Beck's Depression Inventory II (BDI-II; Hautzinger et al., 2006) was assessed pre- and post-intervention. To assess the change in depressive symptoms over treatment, we calculated the percentage difference of BDI-II scores between T0 (baseline) and T7 (post treatment).

Statistical Modelling

All three interventions (CBT, ST and IST) were found to be clinically useful in the treatment of depression (Kopf-Beck et al., 2024). The primary analysis of the OPTIMA study found no significant impact of concomitant care conditions on the differences between intervention conditions. Although, this is not the focus of this study, to account for differences in concreteness dependent on the treatment arms, we controlled for treatment group as a predictor of concreteness in our analysis of hypothesis 2. According to the recommendations by Lee and Hong (2021), three-level models require a minimum of 50 units per group for fixed effects and at least 100 units for random effects, criteria which our data do not meet. Thus, in the subsequent analyses, the three treatment arms are combined.

To test the first hypothesis, that momentary RNT concreteness explains variance in the prediction of momentary depressive symptoms beyond momentary RNT, two-level multi-level modelling (MLM) was used to account for the nested structure of the data (i.e. multiple measurement time points within each patient). The MLM was specified as follows:

$$\begin{aligned} MomDep_{ij} = {} & \beta_0 Int_i + \beta_1 MomRNT_{ij} + \beta_2 MomCon_{ij} + \beta_3 Time_{ij} \\ & + \beta_4\left(MomRNT_{ij} * Time_{ij}\right) + \beta_5\left(MomCon_{ij} * Time_{ij}\right) \\ & + \beta_6\left(MomCon_{ij} * MomRNT_{ij}\right) + \varepsilon_{ij} \\ \beta_0 Int_i = {} & \gamma_{00} + u_{0i} \end{aligned} \quad (1)$$

where $MomDep_{ij}$ reflects momentary depressive symptoms on prompt level of the i-th participant at timepoint j. $MomDep_{ij}$ is predicted by the participant's momentary RNT at the same timepoint ($MomRNT_{ij}$), momentary concreteness at the same timepoint ($MomCon_{ij}$), the time ($Time_{ij}$; ranging from 0 to 167), the interactions of $MomRNT_{ij}$ and

$Time_{ij}$, $MomCon_{ij}$ and $Time_{ij}$, as well as $MomCon_{ij}$ and $MomRNT_{ij}$. Individuals were allowed to differ randomly in baseline levels of $MomDep_{ij}$, denoted by the fixed intercept γ_{00} and the random deviation by participant u_{0i} (i.e., random intercept). The remaining factors in the models were fixed effects. The residual accounting for the unexplained variance in momentary depressive symptoms by the *i*-th participant at timepoint *j* is denoted by ε_{ij}. The fit of the model was then compared to a simpler model without concreteness ($MomCon_{ij}$, $MomCon_{ij} * Time_{ij}$) with ANOVA-model-comparisons.

Furthermore, we examined the association between the concreteness of RNT and momentary RNT with a Pearson correlation between the two measures.

In examination of the second hypothesis, that momentary RNT concreteness increases over the course of psychotherapy, we used two-level MLM to test the following model:

$$MomCon_{ij} = \beta_0 Int_i + \beta_1 Time_{ij} + \beta_2 BDI\ impr_i + \beta_3 (Time_{ij} * BDI\ impr_i) + \varepsilon_{ij} \quad (2)$$

$$\beta_0 Int_i = \gamma_{00} + u_{0i}$$

where $MomCon_{ij}$ is predicted by the $Time_{ij}$ and the percentage change of the BDI-II from pre- to post-treatment ($BDI\ impr_i$) as well as the interaction between $Time_{ij}$ and $BDI\ impr_i$. Individuals were allowed to differ randomly in baseline levels of $MomCon_{ij}$, denoted by the fixed intercept γ_{00} and the random deviation by participant u_{0i}. The remaining factors in the model were fixed effects.

To examine whether the concreteness changes differently in the three treatment arms, we added the treatment group as a control variable to the analysis. Additionally, we investigated its interaction with $Time_{ij}$ and $BDI\ impr_i$. As we did not find a significant effect of the group on concreteness of RNT, the simpler model is reported in the following analyses. For transparency and because of the value of this result for further investigations, we report these results in Table C.3 in the Supplementary Material.

To examine the exploratory hypothesis regarding the temporal precedence between momentary RNT concreteness and momentary depressive symptoms, we tested the following two-level MLM:

$$MomDep_{ij} = \beta_0 Int_i + \beta_1 MomCon_{i(j-1)} + \beta_2 MomDep_{i(j-1)} + \varepsilon_{ij} \quad (3.1)$$

$$\beta_0 Int_i = \gamma_{00} + u_{0i}$$

where $MomDep_{ij}$ is predicted by the participant's $MomCon_{ij}$ at the previous timepoint $(j-1)$ and the autoregressive control parameter $MomDep_{ij}$ at the previous timepoint $(j-1)$.

We furthermore tested the reverse relationship, whether $MomCon_{ij}$is predicted by $MomDep_{ij}$ at the previous timepoint $(j-1)$, controlling for the autoregressive parameter of $MomCon_{ij}$ at the previous timepoint $(j-1)$. Individuals were allowed to differ randomly in baseline levels of the respective dependent variable ($MomDep_{ij}$ or $MomCon_{ij}$), denoted by the fixed intercept γ_{00} and the random deviation by participant u_{0i}. The remaining factors in the models were fixed effects.

$$MomCon_{ij} = \beta_0 Int_i + \beta_1 MomDep_{i(j-1)} + \beta_2 MomCon_{i(j-1)} + \varepsilon_{ij} \quad (3.2)$$
$$\beta_0 Int_i = \gamma_{00} + u_{0i}$$

To determine whether the inclusion of control variables age and gender significantly enhanced the explained variance, for all hypothesis, more complex models including age and gender were compared to the simpler models. The results indicated that age and gender did not contribute significantly to the variance explained by the models. Therefore, the simpler models are reported in the following analyses. Details about the model comparisons can be found in Table C.4 in the Supplementary Material. All statistical preprocessing and analyses were performed with R-Statistics (R Core Team, 2021). Calculations were made with the package "lme4" (Bates, Maechler, Bolker, & Walker, 2015) and results were plotted with "jtools" (Long, 2022) and "ggplot" (Wickham, 2016).

Results

Sample description

The study was conducted from August 2019 to December 2020. The mean age of the final sample was on average 41 years (*SD*=12.30, range=21 - 71), 54.55% were female and 84.42% were German. Thirty-one (40.26%) patients were allocated to the CBT, 23 (29.87%) to the ST and 23 (29.87%) to the IST intervention condition. At baseline, patients showed severe levels of depression on average (BDI-II: *M*=32.49, *SD*=8.26), which decreased to mild levels by the end of the treatment (BDI-II: *M*=17.04, *SD*=8.92). Further descriptives of the sample can be found in Table 4.1.

Table 4.1

Descriptive Statistics of the Sample (n=77)

Characteristic	***N***	**%**
Gender (female)	42	54.55%
Nationality (German)	65	84.42%
School graduation (Qualification for University entrance)	38	49.35%
Income		
Low income (up to 1500 EUR)	30	38.96%
Middle income (1500 - 4000 EUR)	30	38.96%
High income (more than 4000 EUR)	12	15.58%
not specified	5	6.49%
	M	***SD***
Age (years)	41	12.30
EMA response rate in %	54.57	26.78
Baseline depression severity (BDI-II)	32.49	8.26
Baseline Momentary depressive symptoms	1.69	0.73
Baseline Momentary RNT	1.72	0.88
Baseline Momentary Concreteness	2.47	0.65

Note. *n* = 77; EMA: Ecological Momentary Assessment; BDI-II: Beck's Depression Inventory II. Baseline = mean of the baseline week 0. The variable depressive symptoms ranged from 0-4, the variable momentary RNT ranged from 0-4 and the variable momentary concreteness ranged from 1-5.

Hypothesis 1: Prediction of momentary depressive symptoms by the concreteness of momentary RNT

To test whether the concreteness of momentary RNT is associated with the level of momentary depressive symptoms assessed at the same timepoint, we ran two MLMs with momentary depressive symptoms as the dependent variable and (A) time and momentary RNT as well as their interaction as predictors vs. (B) time, momentary RNT and concreteness as well as their interactions as predictors (Table 4.2). In Model A, we found a significant effect of momentary RNT on momentary depressive symptoms, with higher momentary RNT

predicting higher momentary depressive symptoms, *B*=0.45114, 95% CI=[0.41868 – 0.48360], $p < .001$. Furthermore, we found a significant effect of time, *B*=-0.00166, 95% CI=[-0.00229 – -0.00104], $p < .001$, showing earlier timepoints in treatment predicting higher momentary depressive symptoms, but no interaction effect.

Model B shows that lower momentary concreteness significantly predicts higher momentary depressive symptoms, *B*=-0.12257, 95% CI=[-0.18471 – -0.06043], $p < .001$ and model comparison-analysis using an ANOVA test revealed that the addition of the predictor concreteness and its interactions improved the model fit significantly as indicated by Likelihood Ratio=93.23, $p < .001$ and a smaller AIC and BIC for Model B (AIC: 9411.6, BIC: 9476.5) in comparison to Model A (AIC: 9496.8, BIC: 9535.8).

The results of the correlation between momentary RNT and the concreteness of RNT show small correlations of *r*=-0.29 (*p*-value < .001).

Table 4.2

MLM's of Momentary depressive symptoms Predicted by Momentary RNT Versus Momentary RNT and the Concreteness of Momentary RNT (MomCon)

IDV/predictors	Estimates	SE	t	*p*	95% CI
Momentary Depressive Symptoms (Model A)					
Intercept	0.98730	0.08	11.88	<.001	[0.74763 – 1.04646]
Time	-0.00166	<0.01	-5.2	<.001	[-0.00229 – -0.00104]
MomRNT	0.45114	0.02	27.25	<.001	[0.41868 – 0.48360]
MomRNT*Time	0.00021	<0.01	1.21	.227	[-0.00013 – 0.00056]
Momentary Depressive Symptoms (Model B)					
Intercept	1.22117	0.11	10.74	<.001	[0.99757 – 1.44428]
Time	−0.00218	<0.01	-2.27	.023	[-0.00405 – -0.00030]
MomRNT	0.43090	0.04	10.22	<.001	[0.34827 – 0.51353]
MomRNT*Time	0.00035	<0.01	0.74	.458	[-0.00058 – 0.00129]
MomCon	-0.12257	0.03	-3.87	<.001	[-0.18471 – -0.06043]
MomCon*Time	0.00017	<0.01	0.52	.603	[-0.00048 – 0.00082]
MomRNT*MomCon	0.00202	0.02	0.13	.897	[-0.02838 – 0.03242]
MomRNT*MomCon*Time	-0.00006	<0.01	-0.33	.738	[-0.00042 – 0.00029]

Note. *n*=77; Model A: ICC=.47; Marginal R^2=.308, Conditional R^2=.634. Model B: ICC=.47; Marginal R^2=.330, Conditional R^2=.644; Estimates = unstandardised regression coefficients; MomCon: Concreteness of momentary RNT; MomRNT: Momentary Repetitive Negative Thinking.

Hypothesis 2: Prediction of Concreteness of RNT by time and BDI-II improvement

A MLM was conducted to test whether the concreteness of RNT increased throughout psychological treatment (Table 4.3). In this model we found that on average, earlier treatment timepoints significantly predict higher concreteness levels, *B*=-0.00263, 95% CI=[-0.00371 – -0.00155], $p < .001$. However, we found that this association depended on the BDI-II improvement as indicated by a significant interaction between time and BDI-II improvement, *B*=0.00005, 95% CI=[0.00003 – 0.00007], $p < .001$ (Table 4.3), with above-average BDI-II improvement predicting increases in concreteness over time and below-average BDI-II improvement predicting decreases in concreteness (Figure 4.2).

Table 4.3

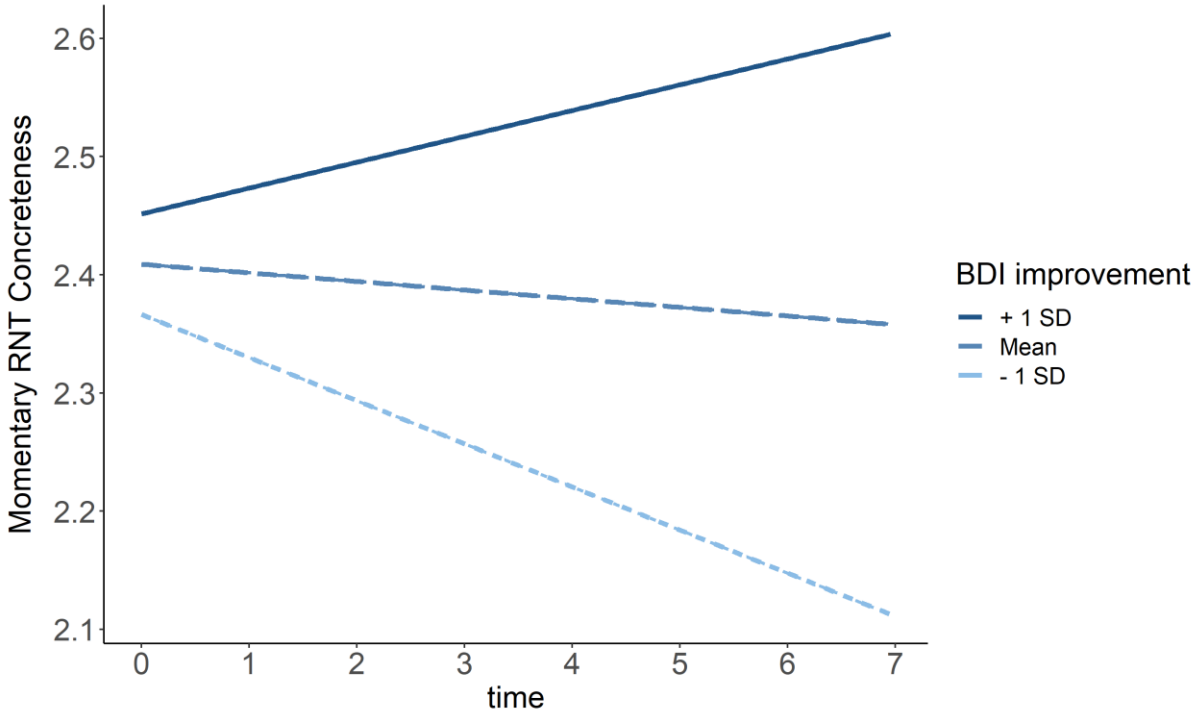
MLM of Concreteness of Momentary Repetitive Negative Thinking predicted by Time and Improvement of Depression Severity (BDI-II) from baseline to the end of the intervention.

IDV/predictors	Estimates	SE	t	*p*	95% CI
Intercept	2.32754	0.16	14.86	<.001	[2.01616 – 2.63728]
Time	-0.00263	<0.01	-4.77	<.001	[-0.00371 – -0.00155]
BDI-II Impr	0.00175	<0.01	0.6	.553	[-0.00408 –0.00760]
BDI-II Impr*Time	0.00005	<0.01	4.73	<.001	[0.00003 – 0.00007]

Note. *n*=77; ICC=.35; Marginal R^2=.024, Conditional R^2=.362; Estimates = unstandardised regression coefficients; BDI-II Impr: BDI-II Improvement; BDI-II: Beck's Depression Inventory.

Figure 4.2

Linear Regressions of Momentary RNT Concreteness over the course of the seven-week intervention period



Note. n = 77; time: treatment weeks; The figure illustrates the linear modelling of change in momentary RNT concreteness measured with ecological momentary assessment over the course of the seven-week intervention period separately for people with mean BDI-II improvement, 1 SD above average and 1 SD below average BDI-II improvement from baseline to the end of the intervention. RNT: Repetitive Negative Thinking; BDI-II: Beck's Depression Inventory (BDI-II).

Exploratory research question: Temporal precedence between momentary concreteness and momentary depressive symptoms.

Finally, we investigated the temporal precedence between momentary concreteness and momentary depressive symptoms in a time-lagged MLM (Table 4.4). We found a significant prediction of concreteness by momentary depressive symptoms at the previous timepoint, B=-0.06, 95% CI=[-0.09 to -0.02], p=.001, whereas the prediction of momentary depressive symptoms by concreteness at the previous timepoint was not significant.

Table 4.4

MLM's of Momentary Depressive Symptoms (MomDep) and Momentary Concreteness (MomCon) predicted by the respective other variable at the previous timepoint.

IDV/predictors	**Estimates**	**SE**	**t**	***p***	**95% CI**
Momentary Depressive Symptoms					
Intercept	1.03	0.08	13.04	.001	[0.87 – 1.18]
MomCon at t-1	-0.02	0.02	-1.32	.186	[-0.05 – 0.01]
MomDep at t-1	0.35	0.02	21.02	<.001	[0.32 – 0.38]
Concreteness of MomRNT					
Intercept	2.20	0.09	25.70	<.001	[2.03 – 2.36]
MomDep at t-1	-0.06	0.02	-3.28	.001	[-0.09 – -0.02]
MomCon at t-1	0.13	0.02	7.48	<.001	[0.10 – 0.17]

Note. *n*=70; Model A: ICC=.30; Marginal R^2=.171, Conditional R^2=.419. Model B: ICC=.30; Marginal R^2=.031, Conditional R^2=.322; Estimates = unstandardised regression coefficients; MomRNT: Momentary Repetitive Negative Thinking; *MomDep: Momentary Depressive Symptoms; MomCon: Concreteness of momentary RNT.*

Discussion

This study was the first to investigate momentary levels of RNT concreteness and their predictive value for depression over the course of psychological interventions. Within this framework, we investigated three primary objectives: (a) whether momentary RNT concreteness explains variance in the prediction of momentary depressive symptoms beyond momentary RNT, (b) the change of momentary RNT concreteness over the course of treatment, and (c) the temporal relationship between momentary RNT concreteness and momentary depressive symptoms.

Prediction of momentary depressive symptoms by the concreteness of momentary RNT

In support of our first hypothesis, we found that the concreteness of RNT significantly predicted momentary levels of depression and accounted for additional variance beyond what was explained by momentary RNT alone. This implies that a more concrete style of RNT is associated with fewer depressive symptoms at the same timepoint and that the concreteness of

RNT delivers an additional value in explaining symptoms of depression. Interestingly, we found no interaction effect, neither between RNT and time, nor between concreteness and time, suggesting that the dynamics between momentary depressive symptoms and momentary RNT or concreteness remain constant over the course of treatment. We neither found an interaction effect between momentary RNT and concreteness in predicting depressive symptoms. This suggests that both, the concreteness of RNT and momentary RNT predict depressive symptoms independently of each other. However, as in the RNT literature there is no clear cut between what is a defining (Ehring et al., 2011) vs. associated feature of RNT, we additionally examined the relationship between RNT and the concreteness of RNT using Pearson correlations. The significant but small correlations we found support the view of concreteness as a distinct associated component of RNT.

Our results extend previous research reporting associations between concreteness and depression (Kircanski et al., 2015; Takano & Tanno, 2010; Watkins & Moulds, 2007) in three major aspects. First, the results replicate the association between concreteness and depression when concreteness is mapped in a high-frequency EMA-format close to everyday life. In prior studies, concreteness was primarily assessed through instructions to recall own memories (Werner-Seidler & Moulds, 2012), to describe current problems and their possible consequences (e.g. Problem Elaboration Questionnaire, Stöber & Borkovec, 2002), or through self-ratings of concreteness (Kircanski et al., 2015). In contrast to that, EMA data assessed via smartphones is widely accessible, integrates into patients' everyday life and prevents biases due to the recall of memories. The text units assessed in our study were quite similar to the communicative form of text messages or app-based diary entries and could therefore contribute to the assessment and transfer of therapeutic achievements into everyday life.

Secondly, we replicated previous findings in the context of an inpatient psychiatric setting. Our results demonstrate the stability of the association between concreteness and depression over the course of psychological interventions, while previous studies have mainly investigated concreteness in an experimental setting (Werner-Seidler & Moulds, 2012) or over shorter time-periods and without treatment (Takano & Tanno, 2010).

Third, our findings underline the importance of concreteness as an additional predictor of momentary depressive symptoms beyond the process of RNT, illuminating the "how". Previous research examining the role of concreteness within the relationship between rumination and depression (Starr et al., 2017; Takano & Tanno, 2010) is thus complemented

by our findings, contributing to a deeper understanding of depression in patients with clinically diagnosed depression.

Prediction of Concreteness of RNT by time and BDI-II improvement

While investigating the change of concreteness over the course of treatment, contrary to our second hypothesis, we found that on average, patients' concreteness levels decreased slightly throughout treatment. However, this association interacted with patients' BDI-II-improvement, indicating that in patients with an above-average BDI-II improvement, concreteness increased slightly over time, while patients with a below-average BDI-II improvement demonstrated a decrease in concreteness. A possible reason explaining these results could be that at the start of treatment, some patients firstly confront themselves with their problems and therapy goals, which might exacerbate concrete thinking at the beginning of therapy. Throughout treatment, patients who better respond to therapy, indicated by an above-average improvement in depression severity, learn to think concretely, whereas patients who improve less in depression severity do not. It may also be the case that the concreteness of RNT requires a longer period to improve during depression treatment compared to changes in depression severity. However, distinguishing the change of concreteness from methodological aspects such as change sensitivity of the measurement tool used can be challenging. Concludingly, rumination has already been described as a therapy-interfering behaviour (Watkins & Roberts, 2020). Our results implicate that it might not be rumination in general but specifically abstract rumination which may hinder the effectiveness of psychological interventions as a maladaptive emotion-avoidant coping strategy and thus improvement in depression.

The results parallel and extend findings from Stöber, who observed higher concreteness levels in patients with GAD compared to patients with depression following CBT (Stöber & Borkovec, 2002). To our knowledge, this is the first study that found concreteness increasing throughout therapy in dependency of patients' depression improvement, although our interventions did not specifically target concreteness (Watkins & Moberly, 2009). The interventions applied in our study were CBT, which is considered as the gold standard in clinical practice (David, Cristea, & Hofmann, 2018) and other well-established therapies such as ST and IST, that are also increasingly researched and applied for different disorders such as depression (Kopf-Beck et al., 2024). In coherence with the results of our first hypothesis it would therefore be interesting to further investigate whether the

concreteness of RNT is a concomitant phenomenon of depression treatment (i.e. through the reduction of global self-judgements, cognitive restructuring, behavioural activation or emotion regulation strategies) and whether this differentiates between different forms of therapy.

Temporal precedence between momentary concreteness and momentary depressive symptoms.

Beyond concurrent associations, we explored the temporal association between improvements in momentary depressive symptoms following concreteness levels or vice versa. Our results indicate that change in momentary concreteness follows change in momentary depressive symptoms and abstract thinking may therefore be a consequence of depression symptoms such as low mood.

Previous studies only investigated the temporal precedence between rumination and depressive symptoms, showing that rumination on the previous timepoint predicts following negative affect (Kircanski et al., 2018). Our results suggest that this relationship is different for the concreteness of RNT in contrast to rumination, as we did not find concreteness to predict momentary depressive symptoms.

Next to the assessment of concreteness with trained raters, an important difference between our study and the study of Kircanski (2018) is that we applied psychological treatment for depression. Thus, an explanation for the results might be that depression was directly addressed by our treatment and thus broad depression measures change first. It might also be explained by the observed time interval as we only measured short-term-relationships and didn't look out for long-term changes of depression following concreteness improvements. To drive for further insights to the short-term dynamics of concreteness and depression, it would be interesting to test this relationship when concreteness is targeted by the treatment, for example concreteness trainings (Funk et al., 2024) or rfCBT (Roberts et al., 2021), where the opposite relationship would be expected.

Besides that, the results could be explained through the following processes: Due to a bias in negative information processing and an attentional bias towards negative stimuli (Everaert, Podina, & Koster, 2017), depressed individuals could initially experience depressive symptoms caused by other mechanisms than rumination (e.g. sleep deprivation or neurochemical changes). Negative mood in turn could lead to more maladaptive, abstract rumination as a coping strategy. Moreover, the HEXAGON-model, which explains

underlying mechanisms of rumination ("H-EX-A-GO-N: Habit development, EXecutive control, Abstract processing, GOal discrepancies, Negative bias", Watkins & Roberts, 2020, p. 1), suggests that rumination as well as abstract processing have habitual characteristics, that could be reactivated through negative mood (described as H-A-N-interaction in the HEXAGON model). Changing a habit is often difficult and interventions like psychoeducation, cognitive restructuring, as well as changing beliefs and attitudes are not expected to directly address the habitual quality of rumination (Watkins & Roberts, 2020). Thus, habit rumination and abstract processing may need more time than seven weeks to decrease and to be replaced by concrete thinking as an adaptive coping strategy. The association between concreteness predicting depression on the subsequent timepoint and throughout psychological treatment for depression should thus be further investigated in long-term interventions addressing habitual characteristics of RNT.

Methodological considerations

Next to the clinically relevant findings on the association between concreteness levels of daily RNT and depressive symptoms, the current study delivers further methodological insights. To the best of our knowledge, this is the first study implementing Stöber's Concreteness Scale on short units of text that vary widely in length and language and were not assessed with the goal of measuring concreteness. By extending the Concreteness Scale to the development of a decision aid, it is also capable to produce ratings which ensured a sufficient interrater reliability of Krippendorff's Alpha=0.74. This can be assumed as satisfactory since the data set does not provide good conditions for high interrater reliability. In addition, the rating process of concreteness can in general be considered as an interpretative performance which clearly exceeds a purely observational performance. To compensate for this, regular quality checks were in place during both the training and evaluation phase. Raters were graduated psychologists and therefore experts regarding the clinical context of the survey and thus particularly well suited for psychological ratings. It would be interesting to further apply the decision aid in other intervention contexts such as concreteness trainings and rfCBT as well as to different text forms and lengths.

Limitations

Despite the support of our hypotheses, the results of our study require the discussion of several limitations. First, the substantial proportion of missing data and excluded text units (e.g. text units of "no thoughts") limits our conclusions to patients who were aware of their

RNT and willing to write down their thoughts in a detailed enough way for them to be rated for concreteness. Also, nonsense expressions were not analysed although they may indicate an abstract thinking style and difficulties or high effort to think concretely enough to write down any kind of thought. By the exclusion of positive valenced text units due to not answering the question of the item correctly we could have also excluded thoughts of persons that already improved in concrete thinking and experienced less depressed mood. However, since this type of text accounts for only 60 excluded units, it is unlikely to be relevant for the present sample. Moreover, our conclusions are limited to RNT thoughts, which means they do not account for situations in which patients had no RNT thoughts. For example, the expression "no thoughts" could indicate a concrete style but was not analysed due to low expressiveness for our hypothesis. Furthermore, the exclusion of missing post-treatment BDI-II data and the requirement for participants to attend more than 78% of treatment sessions, may limit the generalizability of the results. While there are various methods for handling missing data (e.g., multiple imputation) and missing treatment dose (e.g., intent-to-treat), we followed the approach specified in the study protocol (for details see Kopf-Beck, 2020) in line with other analyses in this study framework (Tamm et al., 2024).

As discussed, another limitation is that we looked only at time precedence on prompt level but not at long-term time intervals such as days, weeks or even longer periods. It is possible that an improvement in concreteness may have long-term effects on the alleviation of depression, which could only be observed over extended time intervals, while low mood directly leads to more abstract rumination or worry.

Furthermore, while all three treatment arms effectively reduced depression and group effects did not significantly predict changes in concreteness over time, different underlying processes may have contributed to these outcomes. However, due to the small sample size, a differentiated comparative analysis of RNT processes across the three treatment arms was not conducted (Lee and Hong, 2021).

As worry and rumination are hypothesised to have a qualitative overlap (Watkins & Moulds, 2007), the present work does not distinguish both forms but considers RNT as a transdiagnostic process influencing the development and maintenance of several emotional disorders like depression (Nolen-Hoeksema & Watkins, 2011). This is supported by the evidence of a single latent factor underlying the influence of RNT on psychopathology (Arditte, Shaw, & Timpano, 2016). However, recent research found differences in

concreteness associated with rumination versus worry (Kircanski et al., 2018). Therefore, further research needs to confirm the generalisability of our results.

Finally, it is important to note that our study was the first utilising the developed decision aid. Validating the inter-rater reliability achieved in our study through a subsequent study would further enhance the quality and robustness of the decision aid.

Future directions

The feasibility of manual concreteness ratings for practical everyday therapy is limited in that both the training of raters and the rating itself require large amounts of time that hardly any therapist can invest. However, assessing the concreteness of RNT with automatic approaches, such as dictionary-based approaches (e.g. the Linguistic Inquiry and Word Count, Pennebaker, Booth, Boyd, & Francis, 2015) is also challenging. These methods rank the concreteness of single words and are primarily based on the criteria of whether or not the meaning of a word can be experienced through the senses (Brysbaert, Warriner, & Kuperman, 2014; Köper & Im Schulte Walde, 2016). For instance, words like "heat" are considered more concrete than "justice" due to their direct sensory associations. However, dictionary-based approaches can struggle with phrases where the context, in which the words are written, plays a crucial role, leading to potential misclassification. For example, "justice" might be ranked abstractly despite concrete contextual meanings, such as when it refers to the fair distribution of four apples to two children. A promising method for creating economical and precise ratings could thus be the use of large language models (LLMs, Stade et al., 2024). In the past years, they were increasingly researched for their contribution to mental health analysis (Lan et al., 2024; Yang et al., 2023). Therefore, it seems promising to investigate whether LLMs can produce valid ratings and to compare these ratings with the manual ratings assessed in our study. As LLMs deliver the advantage of being ecologically accessible and highly scalable, concludingly, an automated feedback tool for the patient and therapist regarding the concreteness of thoughts could be developed. As we are confronted daily with our thoughts about internal and external events, this could be an innovative and effective way to restructure thought patterns and develop new thinking styles. By automatically capturing concreteness, such tools could be used in smart sensing, and as an adjunct to psychotherapy to monitor and support the treatment process. Temporal relationships between concreteness and depression could thus be further explored in the context of concreteness trainings using such tools to further understand the particular dynamics of changes of concreteness and depression over time.

Conclusion

Despite the limitations, the results of the current study underscore the relevance of RNT concreteness for depression. Further research is needed to explore the concreteness of RNT as a mediator between RNT and depression in different forms of therapy, investigating its role as a potential mechanism of change of depression.

5. General Discussion

Summary and further discussion of the findings

This dissertation investigated different opportunities of EMA to support psychotherapy research in depression. Over a period of 17 months (August 2019 - December 2020), an EMA study was conducted within a large RCT investigating the effectiveness of three different psychotherapy approaches. The EMA substudy initially included 106 depressed patients, which is about one-third of the entire RCT sample. All patients were randomized to seven weeks of cognitive behavioral therapy (CBT), schema therapy (ST), or individual supportive therapy (IST), including two group sessions of 100 minutes each and two individual sessions of 50 minutes each per week. The treatments were delivered in an inpatient or dayclinic setting in parallel with standard psychiatric treatments, including pharmacotherapy and additional treatments such as ergotherapy or case management (Kopf-Beck et al., 2024). In addition to a comprehensive test battery of the RCT (for details see the OTIMA study protocol: Kopf-Beck et al., 2020), patients of the EMA substudy provided daily momentary self-reports of depressive symptoms and RNT three times a day throughout their entire intervention period.

Based on the collected EMA data, a total of eight different research questions were addressed. The results were reported in three distinct studies and research articles, providing empirical insights into the potential of EMA to improve our assessment, treatment, and understanding of depression.

Study I

Study I examined the comparability of EMA and WQA in assessing intervention effects and predicting change clincian-rated global functioning. The study successfully replicated the findings of a previous study (Moore et al., 2016) in that EMA is more sensitive than questionnaires in detecting differences between intervention effects. Specifically, EMA revealed that ST was more effective in reducing RNT than the other two intervention groups, CBT and IST. In terms of predicting changes in clinician-rated global functioning, WQA had greater predictive accuracy, although EMA-assessed changes in depressive symptoms also yielded significant predictions. In addition, we observed notable differences in time effects (slopes) between the two assessment techniques: WQA scores showed a steeper decline over time and were more extreme, with higher baseline and lower post-intervention values, compared to EMA scores.

As the EMA slopes were smaller than the WQA slopes, we propose that the higher sensitivity of EMA to detect differences between intervention effects is not simply attributable to its greater sensitivity to change. Rather, we suggest that EMA may more reliably assess change in depression by avoiding recall biases (Colombo, Suso-Ribera et al., 2019) to which retrospective questionnaires are prone to, allowing EMA to detect small differences between intervention effects with greater statistical power. Interestingly, higher amplitudes of change in questionnaires than EMA were also found in a pharmacological study examining treatment effects on quality of life with EMA versus questionnaires (Barge-Schaapveld & Nicolson, 2002).

These findings suggest that retrospective questionnaires may systematically overestimate the amplitude between baseline and post-treatment symptom severity, i.e., the intervention effect. While recall biases occur in all humans, they manifest more strongly in individuals with depression, particularly with regard to the retrospective overestimation of negative affect (Colombo, Suso-Ribera et al., 2019; Colombo et al., 2020). That is, depressed patients may overestimate the severity of their symptoms, particularly early in treatment when their depression is still quite severe, compared to later in treatment when both their symptoms and cognitive biases may have improved with therapy. The recall biases inherent in retrospective questionnaires may stem from expectancy effects (e.g., expectations about treatment outcomes; Zetsche et al., 2019), reappraisal (e.g., re-evaluating difficult experiences after they have passed; Levine et al., 2012), or mood-congruent memory effects (e.g., better recall of experiences consistent with one's current mood; for a review, see: Ebner-Priemer & Trull, 2009a). Importantly, some of these biases, such as reappraisal, are explicitly addressed by psychotherapeutic techniques like cognitive restructuring, a core component of CBT (Beck, 2021).

In addition, the finding that EMA significantly predicted changes in global functioning as assessed by clinical interviews supports its external validity. An explanation for the superior prediction accuracy of WQA may be the shared retrospectivity of clinical interviews and questionnaires. As both are retrospective point-assessments, they may be similarly affected by recall bias, alongside additional biases that can occur in interviews, such as interviewee social desirability bias or biases arising from the clinician’s side, such as differing expectations of the clinician at baseline and post-treatment.

In interpreting the results of Study I, however, it is crucial to acknowledge that further research is needed to systematically examine how recall bias evolves over the course of

treatment (Ebner-Priemer & Trull, 2009b), and to explore if EMA's mitigation of recall bias drives its increased sensitivity for detecting intervention effects. For now, this relationship remains hypothetical. The EMA design used in our study may also have introduced biases, such as inaccurate monotonous responses due to participant fatigue with the repetitive assessments. Therefore, future studies should also develop improved EMA designs that include techniques to ensure the reliability and conscientiousness of participants responses.

The study highlights the unique strengths and limitations of EMA, WQA, and clinical interviews, suggesting that the combination of all three assessment techniques may enhance the validity and reliability of clinical assessments in clinical trials.

Study II

Study II explored how early treatment response prediction could be realized in psychological therapies by focusing on the monitoring of early improvements. Given that EMA integrates into patients' daily lives and enables frequent, real-time assessments of change (Colombo, Fernández-Álvarez et al., 2019; Trull & Ebner-Priemer, 2009b), it holds promise as a tool for monitoring early therapeutic progress (Li et al., 2023). However, despite strong evidence supporting early improvement as a robust predictor of treatment response in psychological interventions (Beard & Delgadillo, 2019; Li et al., 2023), there remains a lack of consensus regarding the optimal timing and rate of change that serves as the best predictor of early treatment response, even for questionnaire assessments (Beard & Delgadillo, 2019).

The results of Study II show that both EMA and WQA of early improvement significantly predict patient response (versus non-response) to the full seven-week interventions. As early as three weeks into the treatment, both assessment techniques provided significant predictions of the final treatment outcomes, with comparable AUC values of 73% (EMA) and 77% (WQA). Nevertheless, the questionnaire predictor revealed a clearer pattern of which change rate serves as the best predictor for clinical implications, which we defined as high ratios between the true negative and false negative rates. The best predictor was a WQA-assessed 10% improvement after four weeks of treatment, resulting in a true negative rate of 22% compared to a false negative rate of 0%.

When interpreting these results, it's important to note that both the 'questionnaire-assessed early improvement' predictor and the dependent variable 'treatment response' were operationalized using the same questionnaire, the BDI-II. As this alignment might be an

advantage for the questionnaire predictor, it is surprising that EMA predictions were generally comparably accurate. This potentially “unfair” comparison may explain why the prediction pattern of the questionnaire was clearer than that of EMA. To facilitate a fairer comparison between the two predictors - EMA and questionnaire-assessed early improvement - future studies should employ a more “neutral” dependent variable that does not overlap with one of the predictors, such as global functioning or quality of life.

Nevertheless, the results of Study II highlight the potential of both EMA and questionnaires as reliable techniques for monitoring patient improvement for the purpose of early treatment response prediction. When the change rate cutoff was set at a minimum of 10%, the questionnaire approach yielded a very low false-negative rate of just 4% after three weeks and even 0% after four weeks. This underscores the predictive value of early improvements and suggests that the absence of ≥ 10% improvement, assessed three to four weeks after treatment initiation with a validated questionnaire, is a reliable predictor of treatment non-response. Accordingly, it could inform decisions about modifying ongoing psychological interventions that are meant to last for seven weeks and are comparable with the treatments applied in our study.

Clinical trial designs that focus on the development of personalized therapies and that could use our early improvement predictor are adaptive trial designs such as the sequential, multiple assignment, randomized trial design (SMART; Kidwell & Almirall, 2023) and the leapfrog design (Blackwell et al., 2019). In SMART trials, participants are randomized at two or more stages, with each subsequent randomization and the available treatment options depending on the individuals‘ response to the previously randomized condition. However, SMART trials empirically construct adaptive treatments, but do not provide definitive evidence for the effectiveness of these interventions. Subsequently, it is required to evaluate the effectiveness of these interventions through confirmatory randomized trials (Kidwell & Almirall, 2023). While SMART trials focus on adapting treatments for individual participants, adaptive rolling designs like leapfrog adapt treatment arms across the entire sample based on ongoing results. Specifically, they continuously monitor the effectiveness of simultaneously investigated treatment arms, and based on sequential Bayesian analyses, poorly performing arms are replaced or modified. In the leapfrog design, early treatment response prediction through early improvement monitoring could be utilized to rapidly evaluate the effectiveness of specific treatment arms and foster their adaptation to optimize effectiveness (Blackwell et al., 2019). However, a current limitation is the lack of systematic research on how quickly (or

slowly) individual depressive symptoms and their interrelated dynamics change. The concept of treatment response focuses on short-term effects hat manifest by the end of treatment. From both a patient and economic perspective, however, post-intervention trajectories are of greater importance. To account for long-term effects, future studies should therefore explore the predictive value of early improvements on follow-up outcomes such as retained remission and relapse and recurrence rates.

The findings of Study II also indicate that targeting a high ratio between the true-negative and false-negative rate, as demonstrated in our study, may be more meaningful for clinical implications than focusing solely on the overall accuracy of predictors (i.e., the Youden index), as applied by other studies (e.g., Crits-Christoph et al., 2001).

Study III

Study III utilized EMA to explore the temporal relationships between momentary levels of depressive symptoms and RNT (Ehring & Watkins, 2008; Rosenkranz et al., 2020), as well as their changes over the course of therapy. Unlike previous EMA studies on RNT, we focused specifically on the concreteness (Stöber & Borkovec, 2002) of RNT by having patients journal their RNT thoughts three times daily, which were then rated for concreteness by trained external raters. Content features of RNT, such as concreteness, are much less studied on a momentary level, most likely due to the complexity of their assessment. To assess valid ratings of concreteness levels (rather than relying on patients' self-perceived concreteness), patients must journal their thoughts, which are then rated by trained external raters (Stöber & Borkovec, 2002; Wahl et al., 2019). In contrast to the assessment of Likert-scaled self-ratings, this procedure is much more time-consuming for both patients and investigators. Therefore, it was also an interesting question whether this elaborated assessment significantly enhances our understanding of RNT and its dynamic in depression, or if focusing solely on the occurrence of RNT, assessed through Likert-scaled items (Rosenkranz et al., 2020), explains the same variance.

Study III revealed that a model that included both RNT and concreteness accounted for significantly more variance in depression than a model with RNT alone. Surprisingly, RNT concreteness generally decreased over the course of therapy. Specifically, the change in RNT concreteness interacted with changes in depression severity: in patients with above-average improvement in depression, RNT concreteness slightly increased, while in those with average or below-average improvement, it decreased. Moreover, we found that higher levels

of momentary depressive symptoms predicted subsequent decreases in RNT concreteness, but not vice versa, and this dynamic remained stable over the course of the interventions.

The results of Study III indicate that patients' momentary depressive symptoms are uniquely associated with their experience of RNT as a process, as well as the concreteness of RNT content. To draw valid conclusions from the finding that RNT concreteness changes throughout treatment in relation to patients' depression severity, future studies should conduct mediation analyses (Baron & Kenny, 1986) to explore whether momentary RNT concreteness serves as a mechanism of change in psychological treatments for depression. The temporal dynamic between momentary RNT concreteness and depressive symptoms is surprising, given that previous studies found mutual effects for momentary rumination, i.e., higher levels of negative affect predict increases in subsequent rumination and vice versa (Moberly & Watkins, 2008).

However, the absence of the reverse effect for concreteness, i.e., higher levels of concreteness did not predict subsequent improvements in depressive symptoms in our study, does not disclaim the general existence of this effect. In our study we investigated the temporal dynamic between the concreteness of momentary RNT and momentary depressive symptoms in time frames of approximately half a day from one prompt to the next. However, it is proposed that concrete RNT (as opposed to abstract RNT) is adaptive by focusing on specific, tangible details of a situation, facilitating problem-solving (Watkins & Moulds, 2005a; Watkins, 2008). As a result, the effect of concreteness on depression may be delayed or emerge over longer periods of time, such as days or weeks, or at specific moments, such as when concrete processing leads to the resolution of a problem (Watkins & Moulds, 2005a; Watkins, 2008). An approach to investigate this effect could involve having patients journal their specific problems and then using EMA to monitor how concretely they think about these problems, as well as to capture the moment when a specific problem is resolved. When controlling for the severeness of the problems and mood is tracked simultaneously, it could be examined whether more concrete RNT facilitates faster problem resolutions in daily life, and whether this, in turn, is associated with faster reductions in depressive symptoms.

These new findings from Study III underscore the potential of EMA in enhancing our understanding of highly fluctuating processes in depression, such as RNT. They demonstrate that examining content features like concreteness, in addition to process features of RNT, deepens our understanding of both RNT and depression. The developed decision aid that extends Stöber's concreteness scale (Stöber & Borkovec, 2002) is now available for future

studies. This decision aid enabled us to achieve satisfactory interrater reliability when rating the concreteness of text units, despite their heterogeneity in length and brevity.

Methodological limitations and future directions for improving EMA protocols

One of the most intriguing questions arising after research is: How can we improve in the future? To address this, the next section explores how our EMA approach could be refined in future studies, considering both improvements in data quality from a research perspective and strategies to enhance patient satisfaction with the EMA assessment. Although patients in our study reported high satisfaction with the EMA approach (92.73% rated the app as very good, good, or reasonably good), this feedback only reflects the views of those who completed the study. Furthermore, our study achieved a response rate of about 60%, which is comparable to other EMA studies (Colombo, Fernández-Álvarez et al., 2019), but still leaves significant room for improvement.

Dynamic EMA schedules

Even though the sequence of the EMA items in our study was randomized within each prompt, patients received the identical eleven items three times a day. This repetitive assessment might have induced anchoring or fatigue effects, reducing the reliability of patients‘ responses and/or their response rate. Therefore, it is important to question whether it is necessary to collect each symptom at the same high frequency (Trull & Ebner-Priemer, 2020; Silvia et al., 2013). Important considerations in designing EMA schedules are (1) the speed with which variables fluctuate and change under natural and treatment conditions, (2) the underlying research question, and (3) the number of items representing a variable.

Affective symptoms, such as low mood, or cognitive processes like RNT vary highly within and between days (Peeters et al., 2006; Takano & Tanno, 2011; Wirz-Justice, 2008) and there is clear evidence that people fail in their precise recall (Colombo, Suso-Ribera et al., 2019). In contrast, other symptoms of depression, such as psychomotor agitation or inhibition, which we also assessed three times per day, may change slower over time (Snippe et al., 2021). This may also apply to changes in quality of life (Barge-Schaapveld & Nicolson, 2002; Trivedi et al., 2006), negative cognitions such as metacognitive beliefs or dysfunctional attitudes (Beck et al., 1979), and personality traits (Bleidorn et al., 2022). Future research should therefore develop precise EMA schedules by systematically investigating the speed in

which individual symptoms naturally fluctuate and change throughout treatments (Snippe et al., 2021).

Another important consideration when designing EMA schedules is the underlying research question. For assessing change in symptomatology over time, end-of-day assessments, where participants provide one average rating at the end of the day that summarizes their experiences throughout the day, may be sufficient and reduce the EMA frequency and related patient burden. To investigate the temporal dynamics between variables, however, momentary (as opposed to averaged) self-reports are necessary, which may need to be collected at a high frequency to capture the required dynamics in real-time (Trull & Ebner-Priemer, 2020). In this regard, it is important to note that previous studies on the temporal dynamics between affect and RNT have used higher sampling frequencies, such as eight prompts per day (Moberly & Watkins, 2008; Ruscio et al., 2015; Kircanski et al., 2018), compared to the three prompts per day used in our study. Theoretically, EMA schedules could also be tailored to individual fluctuation patterns. For instance, mood could initially be monitored at a high frequency, which is then adapted to a person's observed natural fluctuation rate. However, studying different mood change profiles in depressed patients during therapy, van Genugten et al. (2022) found that patients differed in their mood variability (i.e., the magnitude of mood changes) but not in their emotional inertia (i.e., the speed at which mood shifts over time). This suggests that it may be possible to determine a universal optimal assessment frequency for monitoring mood changes in depressed patients, potentially negating the need for individually tailored frequencies. Nevertheless, it is important to note that van Genugten et al. (2022) did not provide a rationale for the chosen assessment frequency of three times per day, raising the possibility that individual differences in emotional inertia could become apparent at a higher sampling rate.

Finally, when a latent variable of interest is represented by multiple EMA items, two strategies could be employed to reduce participant burden: a) the use of a gate item, and b) a planned missing data design. A gate item is checking the (momentary) occurrence of an event before assessing its specific features. For instance, we could have used a gate item to first assess whether patients were currently engaged in RNT before assessing the four specific features of their momentary RNT (e.g., repetitiveness, intrusiveness, etc.). In this way, patients could have skipped the feature questions whenever they were not currently engaging in RNT, thereby reducing their assessment burden. In our study, we used an item-specific gate-item-like response scale in which patients were first asked whether an item was currently

true before rating the intensity of their agreement on a five-point Likert scale. However, it is important to recognize that in the context of EMA, gate items may introduce potential biases. For example, participants may choose the shortest response path by disagreeing to the gate item if they feel fatigued by the EMA assessments. It cannot be ruled out that this effect also introduced bias into the gate-item-like response-scale used in our study. In a planned missing data design, only a subset of items representing the latent variable is assessed at each prompt, and the items in this subset are systematically rotated from prompt to prompt. The EMA data is then analysed using multilevel structural equation models. This allows researchers to include more items overall without overloading participants at each prompt (Silvia et al., 2013).

Future studies should design their EMA schedules based on empirical evidence about the fluctuations and change rates of variables, as well as based on considerations about the specific research questions and the possibilities of using planned missing data designs. This means each variable (or item) should be assessed at an individual optimal frequency and mode (momentary or aggregated) to maximize reliability while at the same time reducing patient burden (Trull & Ebner-Priemer, 2020). Such an approach could keep a stable data reliability and enable bigger item pools, while increasing patient adherence through a higher variety between EMA prompts.

Validation of EMA approaches

Another important aspect in developing EMA approaches is ensuring their psychometric qualities (Trull & Ebner-Priemer, 2020). In our study, the four items of the momentary RNT score were systematically developed from an item pool and validated in a prior study (Rosenkranz et al., 2020). However, this was not the case for the momentary depression score. The four depression items were created by the authors in consultation with clinicians based on ICD-10 diagnostic criteria for major depression, but they were not systematically developed or validated. Although we assessed and reported good internal reliability for both scores - within and between participants - the momentary depression score lacks the same rigorous psychometric evaluation. To ensure the reliability and validity of EMA, future studies should adopt validated EMA approaches for assessing momentary depressive symptoms. This is also important to enhance the comparability of findings across EMA studies (Ebner-Priemer & Trull, 2009b). A great resource are open-source databases that researchers have begun to develop to improve the transparency, quality, and consistency

of EMA research. In these databases, authors can search for validated EMA approaches or systematically report their own (Hall et al., 2021).

In light of the burden placed on patients through multiple items, a growing debate has emerged around the question of whether it is sufficient to assess psychological constructs in EMA studies using only a single item (Song et al., 2023). Responding accurately to frequent assessments of multiple items requires significant cognitive effort (Krosnick, 1999), and participant fatigue may lead to less thorough responses, resulting in uniform answers both within and across measures and inflated intercorrelations between items. Therefore, although multiple-item measures typically perform better than single items, the added benefit is proposed to be modest in intensive longitudinal designs, supporting the use of single-item measures in such contexts (Song et al., 2023). Future studies should approach this issue systematically. In the case of depression, a highly heterogeneous disorder characterized by symptoms that vary significantly in their nature (e.g., low mood, psychomotor inhibition or agitation, withdrawal; WHO, 2022), relying on a single item to capture the diagnostic construct of depression is unrealistic. However, a single item may still provide valuable information about a specific symptom of interest (Song et al., 2023).

In our study, we used multiple-item measures to assess momentary depressive symptoms and momentary RNT, and all patients were personally instructed to the meanings of all EMA items prior to their first assessment. Considering the aspects discussed, this seems to have been a valid approach. However, it is important to note that our score of momentary depressive symptoms incorporated only four depressive symptoms rather than all eleven symptoms of depression specified in the ICD-10 (WHO, 2022). In the case of RNT, it could be useful to develop approaches with varying levels of detail. In our study, we aimed to capture the dynamics between different RNT processes, which required a multiple-item measure. However, for studies where RNT is not the primary focus and is being assessed alongside several other constructs, a one-item measure might become useful.

Integration of objective data

To meet the complexity of depression research, it is essential to obtain both, high-quality and comprehensive data. This means that the data should be as objective, valid, and reliable as possible. As discussed in the conclusion of Study I, our EMA approach potentially enhanced the reliability of measurements by preventing common biases of retrospective self-reports (Colombo, Suso-Ribera et al., 2019). However, the measurement is not objective, and its

validity can only be determined in relation to other measurement outcomes (e.g., how well it predicts global functioning assessed by clinical interviews). A further leap in data quality could therefore be achieved by incorporating objectively tracked data (Ebner-Priemer & Trull, 2009b). Modern smartphones, especially when combined with wearables, enable the objective monitoring of physiological processes and behavioral patterns, both on-screen and off-screen, including such that correlate with depressive symptoms (Opoku et al., 2021). Meta-analytic findings (Angel et al., 2022) reveal which specific aspects of sleep, physical activity, sociability, locations and phone use exhibit robust correlations with depression. Specifically, depression seems to be robustly correlated with sleep stability, the intensity of physical activity, conversation frequency, time spend at home and phone unlock duration. Additionally, there is meta-analytic evidence that more severe levels of depression are associated with reduced heart rate variability (Koch et al., 2019). In summary, this mean that in our study, for instance, we could have measured patients withdrawal from social activities not only through a subjective EMA item asking, „Are you currently withdrawing from social contacts or activities?“ but also by tracking patients' home stay via GPS and monitoring social interactions through app tracking and/or voice recording. Similarly, we could have tracked patients’ psychomotor agitation or inhibition with wearable devices (Schrijvers et al., 2008), and incorporated variables such as sleep stability and physical activity intensity, which are also trackable with wearables, into our momentary depression score.

To enhance data quality, future studies should therefore consider passive tracking of specific factors. While for affective and cognitive variables, such as mood or RNT, objective data is neither available nor would it be valid, for behavioral and physiological variables passive tracking enables objective measurement and enhances the reliability, frequency and completeness of the data collection. Moreover, passive data tracking reduces assessment burden and may through this enhance patients’ adherence to EMA protocols.

Neutral item wording

From the patient perspective, objective data tracking mitigates not only assessmnt burden but also addresses another risk associated with journaling: a constant self-focus on negative aspects. In our study, patients were asked three times daily, „Which negative thoughts are currently going through your mind repeatedly?“. While many psychotherapeutic interventions aim to shift attention away from negative thoughts towards positive aspects and/or to reduce self-focus (Watkins & Roberts, 2020), frequent journaling of negative aspects might increase ruminative self-focus and/or negative affect (Mor & Winquist, 2002).

Moreover, abstract rumination and worry are proposed to be maladaptive coping strategies with a habitual character (Hjartarson et al., 2021). Especially in individuals prone to abstract RNT as a habitual maladaptive coping strategy, frequent journaling of RNT content may exacerbate RNT and bias its ecological assessment (e.g., “I didn't have any negative thoughts in my mind just now, but now that I'm being asked, I'm again thinking about being a failure.”). For instance, a recent study shows (MacIsaac et al., 2023) that journaling is associated with improvements in psychological wellbeing, but only when it is directed in a positive way and when patients‘ have an average or high disposition to engage in self-reflection. Interestingly, patients‘ disposition to self-reflection did not predict their motivation to journal.

To avoid explicitly directing patients' attention to negative content, future studies should precede the assessment of the RNT content with a gate item, and use a neutral item wording. For instance, a gate item could be ‚Are there currently thoughts running through your mind?’. If individuals answer ‚Yes‘, they could then be asked to describe their thoughts. The principle to use neutral items could also be applied to Likert-scaled items. In our EMA approach, all items except for the mood item were negatively phrased. Instead of asking, „Do you feel like you don’t want to do anything anymore?“ it might be more beneficial to use a neutral phrasing like 'How high is your current interest in pleasurable activities and hobbies?‘. However, it must be considered that even neutral items induce self-focus, which may have positive or negative effects on patients’ momentary mood and/or engagement in RNT. Therefore, it is important to investigate the reactive effects of EMA on patients momentary and long-term wellbeing (van Ballegooijen et al., 2016; Domhardt et al., 2021).

Prevention of negative smartphone effects

Many individuals struggle to limit problematic smartphone use despite being aware of its risks (Busch & McCarthy, 2021). Detrimental effects on mental health through blue light exposure (Heo et al., 2017) and addictive content (Elhai et al., 2017) are well-documented. Future EMA studies should account for these issues. This means, for example, incorporating screen-free times and avoiding blue light exposure before bedtime. In our study, the EMA reminders were already personalized by being scheduled according to each patient's individual wake-up time; however, future studies should ensure that reminders are not sent just before sleep or during times when patients prefer to stay off-screen.

Investigation of patients‘ reactivity to EMA

Besides the potential negative effects of EMA like increasing negative self-focus and supporting problematic smartphone use, EMA can also have positive reactive effects. Many researchers argue that psychological assessments are not just assessments, but interventions in themselves (Poston & Hanson, 2010). While this applies to all forms of psychometric self-reports, it may be especially relevant in the context of high-frequent EMA. Wichers et al. (2011) reviewed the potential advantages of EMA, suggesting that EMA might increase self-awareness, feelings of control over one’s condition, ownership over one’s data, as well as self-management and adherence to psychotherapeutic interventions. When combined with visual feedback on the data collected, EMA may enhance patients’ understanding of their mental health. For instance, in the case of major depression, it could increase patients’ awareness of the variability in their mood, highlighting that their mood is not always low, that periods of low mood are temporary, and providing insights into potential factors associated with mood fluctuations. This may boost patients‘ sense of control and confidence, as well as reduce feelings of guilt associated with their mental health condition. For instance, using EMA to investigate the trajectories of depressed mood, anxious mood, and fatigue in college students, Cranford et al. (2006) found that these symptoms increased before an exam, dropped afterwards, and typically followed a weekly rhythm, with higher average levels on weekdays compared to weekends. In our study, patients had the option to review their previous mood entries, and some patients actively reported to our study team that this helped them recognize the temporary nature of their experience of low mood. In addition, some patients actively requested to continue the assessments after the end of the study, and since this was not restricted, several of them did so.

Despite implementing strategies to minimize reactive EMA effects, such as using neutral items, optimizing EMA schedules, and withholding feedback, it is important to investigate the nature and magnitude of EMA's impact on patients' momentary and long-term wellbeing, as well as to understand under which conditions EMA may make patients "wiser and happier" versus "wiser but sadder" (Domhardt et al., 2021). Factors that may influence these outcomes include the number of items and frequency of the assessments as a source of assessment burden, the wording of items as a source of positive or negative self-attention, and the design of feedback as a source of encouragement or discouragement. In clinical trials, investigating the reactivity of patients to EMA is essential for understanding whether EMA might confound the effects of intentional therapeutic interventions. So far, the reactive effects

of EMA have been largely neglected in EMA research. Therefore, a study protocol by van Ballegooijen et al. (2016), specifically addressing the reactive effects of EMA, seems promising to finally shed light on this important issue.

Further clinical implications and future directions of the utility of EMA in psychotherapy research in depression

The following section delves into further clinical implications and future directions of the utility of EMA in psychotherapy research in depression based on our study findings and the proposed strategies for enhancing EMA approaches.

Improving our clinical assessments with EMA

The findings of our studies highlight the complexity of depression and its reliable monitoring. EMA enables detailed assessments of symptom dynamics (Colombo, Fernández-Álvarez et al., 2019). Studies demonstrating the significant differences between retrospective self-ratings and EMA ratings underline the importance of detailed, real-time assessments (Colombo, Suso-Ribera et al., 2019). This importance is further highlighted by a study from Gratch et al. (2021), which compared the assessment of suicidal thoughts with EMA versus retrospective questionnaires. The study revealed that EMA captures instances of suicidal thoughts that are missed in retrospective reports, and thus proposes that EMA may help to identify at-risk individuals who might otherwise go unnoticed.

At the same time, our study emphasized the crucial role of clinical interviews and traditional questionnaires for clinical assessments. Conducted by clinical experts, structured clinical interviews ensure a comprehensive exploration of symptoms and the valid classification of their clinical significance (Meyer et al., 2001). Therefore, they may be considered a gold standard for diagnosing depression. For instance, a person that overhears a conversation between neighbours might subsequently agree to the question whether having ever heard voices that others couldn't hear. By asking for examples, a clinician could probe further and clarify that the voices the person heard were real and not hallucinatory. In a questionnaire, this opportunity for clarification would be missing. Similarly, many people experience nervousness in specific social situations, such as giving presentations. To assess the clinical significance of such symptoms, it is crucial to explore the range and magnitude of these experiences before considering a diagnosis of social anxiety disorder. The evaluation of clinical symptoms by clinicians is also important to create clarity and trust in the therapeutic

context, especially given patients‘ tendency to self-diagnose (Yıldırım, 2023). Particularly in today’s digital society, misinformation such as about mental health conditions spread rapidly. Although extensive efforts to raise public awareness of mental health problems have fortunately led to better detection of previously un-recognised symptoms, in some people they may also have fostered a problematic over-interpretation of symptoms that lack clinical significance (Foulkes & Andrews, 2023).

However, clinical assessments that rely exclusively on clinical interviews are prone to incomplete understandings (Meyer et al., 2001). Like structured clinical interviews, clinical questionnaires such as the BDI-II, which we used in our study, are normed and have been validated by numerous studies (Hautzinger et al., 2009; Kung et al., 2013). In contrast to structured clinical interviews, however, hey provide highly standardized quantitative information about patients‘ subjective experiences of their symptomatology. Beyond that, questionnaires are much less time- and resource-intensive and their high establishment in clinical research and practice enables the comparability of empirical findings, which is currently a major issue of EMA studies (Ebner-Priemer et al., 2009b; Hall et al., 2021).

These examples illustrate that all three assessment techniques - clinical interviews, questionnaires, and EMA - have their merits, each with distinct strengths. Therefore, it is particularly promising to assign specific roles to each technique and systematically combine them in clinical assessments. Clinical interviews, for instance, might be best used to create detailed profiles of clinically relevant symptoms. For the monitoring of change in these identified symptoms, however, questionnaires and EMA may be more precise. They provide quantitative information, the ease of their administration allows frequent assessments, and they prevent biases that can arise from the patient-clinician dynamic (e.g., expectations of the clinician or social desirability of the patient). Which symptoms should be monitored with questionnaires and which with EMA depends on the natural fluctuation of the investigated symptoms (Fried & Cramer, 2017), the risk of recall bias in the investigated sample and the underlying research question. As outlined before, future studies could develop systematic EMA protocols that assess each item at an individual optimal frequency.

However, decisions about the operationalisation procedures used in a clinical trial must always be made under considerations of the investigated sample and research questions. Retrospective questionnaires aggregate experiences over a past time frame, whereas EMA captures momentary experiences (Colombo, Fernández-Álvarez et al., 2019). Consequently, questionnaires may be well suited to measure trait-like constructs, while EMA is well suited

to capture state-like constructs. Trait und state components however can also coexist in one and the same construct. For instance, a recent study shows that this accounts for RNT (Olatunji et al., 2023). Therefore, it can be assumed that the Perseverative Thinking Questionnaire (PTQ; Ehring & Watkins, 2008), for instance, is a reliable instrument for assessing RNT as a trait, i.e., a person's stable tendency to engage in RNT, whereas the EMA approach used in our study captures RNT states, i.e., a person's momentary engagement in RNT which is only accessible in close proximity to its occurrence.

Another important aspect when distinguishing between retrospective and momentary assessments, is that there are many psychological constructs for those assessment the retrieval of information out of peoples' memories might be particularly necessary and wanted, such as when assessing implicit memories of childhood trauma or metacognitive beliefs about the world and the self, which themselves may not even come to surface in the form of states (Blanke et al., 2022). For other trait variables, it may be theoretically beneficial to develop EMA approaches, especially when their retrospective self-report is unprecise. However, as far as passive tracking of the trait is not possible, the practicality of frequent EMA self-reports may be challenged by the need for long assessment periods and multiple items to capture the stable and complex nature of traits.

The combination of clinical interviews, retrospective self-reports, and EMA may result in more comprehensive and reliable assessments that can help clinicians to better understand and manage the depressive symptoms of their patients and allow for more precise interpretations in clinical trials. Clinical interviews might be best used for diagnostic purposes, EMA for monitoring changes over time, and questionnaires for assessing traits that are reliably reportable retrospectively. The use of smartphones however is well-suited for both high-frequent EMA and low-frequent retrospective assessments, as smartphones are deeply integrated into many people's daily lives (Colombo, Fernández-Álvarez et al., 2019) and allow for easy reminder settings. In the future, it might become the most feasible procedure to assess all kinds of psychometric changes in depressive symptoms with smartphone applications, even when symptoms are assessed at varying frequencies and with different time references, i.e., retrospective versus momentary.

Speeding up the development of personalized modular therapy with EMA

In Study I and II we outlined two strategies through which EMA could support the development of more effective psychological interventions for depression, and specifically

personalized therapies: a) by speeding up the detection of effective interventions through increased measurement reliability, and b) by speeding up the identification of effective interventions through early treatment response prediction. While both strategies aim to accelerate the evaluation process of clinical trials, EMA is also a promising technique to directly enhance the personalization of treatments. The following section will touch on two of such promising strategies: a) the provision of personalized feedback and just-in-time recommendations (Wichers et al., 2011; Colombo, Fernández-Álvarez et al., 2019), and b) the allocation of intervention modules based the analysis of EMA data (Harnas et al., 2021).

Providing personalized real-time feedback and just-in-time recommendations. EMA allows for continuous, personalized real-time feedback, including just-in-time recommendations, which is referred to as ecological momentary intervention (EMI) in the literature. Such feedback holds great potential to enhance the effects of psychotherapeutic interventions: when providing patients with visual feedback and EMIs, patients may become more self-aware of their symptom dynamics (Kauer et al., 2012; Folkersma et al., 2021), gain a deeper understanding of their mental health condition (Myin-Germeys et al., 2018; Folkersma et al., 2021) and develop increased feelings of control and empowerment (Wichers et al., 2011; Colombo, Fernández-Álvarez et al., 2019; Folkersma et al., 2021). This can ultimately reduce depressive symptoms (Kauer et al., 2012; Folkersma et al., 2021), induce behavioral change (Myin-Germeys et al., 2018; Folkersma et al., 2021), support patients in self-managing their condition (Wichers et al., 2011; Folkersma et al., 2021), and increase patients‘ adherence to EMA protocols (Rimpler et al., 2024). Beyond these direct therapeutic effects, EMA feedback may boost patients’ motivation in face-to-face psychotherapy and enhance feelings of transparency in clinical trials. When patients share their data with their clinician, EMA feedback could further support the therapist's work. By providing insights into the patient's well-being between treatment sessions, especially regarding feelings, thoughts, and behaviors that are difficult to report retrospectively (e.g., affective fluctuations and cognitive processes like RNT), or difficult to self-report at all (e.g., behavioral and sleep patterns that can be tracked with wearables), clinicians may gain a clearer understanding of the patient‘s condition, which may further support personalized treatment adaptions (von Klipstein et al., 2020). For instance, Kim et al. (2024) developed a digital application that summarizes patients’ daily experiences from their conversations with an AI-driven chatbot to inform therapists before a treatment session about what happened since the last session. As another example, von Klipstein et al. (2020) and Rimpler et al. (2024) illustrate how the

visualization of EMA data with person-specific network analyses could help therapists and patients in their collaborative exploration of the patient's symptom profile.

Meta-analytic findings show when combined with personalized feedback clinical assessments have robust positive intervention effects of moderate size, especially regarding treatment processes (Poston & Hanson, 2010). Moreover, a meta-analysis investigating the effects of biofeedback shows that feedback on patients' heart rate variability (HRV; i.e., the variation in time between each heartbeat) has positive intervention effects of medium size on the reduction of depressive symptoms (Pizzoli et al., 2021). Greater HRV indicates greater ability of the autonomic nervous system to regulate itself, and reduced HRV is a robust predictor of higher symptom levels of depression (Hartmann et al., 2019), as well as an indicator of higher risk to develop depression (Dell'Acqua et al., 2020).

However, to develop effective feedback tools, it is crucial to explore how mental health states can be communicated to patients in a constructive manner. Simple examples highlight that this is not trivial: constantly informing patients about low mood may be discouraging and demotivating, while only reporting progress and ignoring setbacks may create feelings of intransparency and not being taken seriously. Therefore, research should focus on designing dynamic feedback that remains motivating while avoiding intransparency. In this effort, it is likeli the most promising to collaborate closely with psychotherapists to ensure important aspects of therapeutic feedback, such to paraphrasing, validating feelings, and encouraging problem-solving and adaptive coping strategies. For instance, Hung et al. (2015) developed an EMI (Ecological Momentary Intervention) application designed to assist users in managing negative emotions through emotion regulation strategies like behavioral activation. The application utilizes a machine-learning algorithm that integrates various smartphone data, such as smartphone usage data as an indicator for current mood, as well as behavioral data (e.g., social interaction, physical activity, mindfulness practice, or music engagement), and contextual information (e.g., time, location, and weather) to recommend a situational appropriate emotion regulation strategy.

Allocating digital intervention modules based on EMA data. Another strategy to improve the development of personalized treatments with EMA is to assign treatment modules to patients' needs based on EMA data, either through rule-based or machine learning algorithms. For instance, Harnas et al. (2021) illustrate the personalization of modular CBT for treating cancer-related fatigue in cancer survivors using EMA data. Within a two-week EMA assessment of momentary fatigue levels and potential related factors such as fear of

cancer recurrence, physical activity, and social interactions, the study identified which factors most strongly predicted subsequent levels of fatigue. Each predictor was linked to a specific optional treatment module. The module associated with the strongest predictor for fatigue was then added to the treatment plan, which initially consisted of two mandatory modules. Once all three modules were completed, the process was repeated to further tailor the treatment. However, whether these personalized CBT plans are more effective than one standardized CBT plan for all patients remains to be determined in an RCT.

The fastest way to assess the effectiveness of personalized therapy plans is within a digital framework. The scalability of internet- and mobile-based interventions (IMIs) allows for efficient and intensive data collection, ultimately leading to larger sample sizes and bigger datasets. As Domhardt et al. (2021) reviewed, depression trials of IMIs tend to have significantly larger participant numbers (M = 262, SD = 243) compared to conventional psychotherapy trials for depression (M = 173, SD = 145). The overall effectiveness of IMIs for treating depression is now well-documented (Moshe et al., 2021), but appears to depend on various factors, including the degree of personalization (Hornstein et al., 2023). This is little surprising. In addition to placebo effects, digital evidence-based treatments, like any therapeutic intervention, can only be effective if users actively engage with the treatment modules. This is particularly challenging in depression, as one of the core symptoms of depression is a loss of energy (WHO, 2022). While therapists in face-to-face settings can actively motivate patients toward therapy, unguided digital applications must find creative strategies to capture attention and encourage engagement, competing against the user's screen time on other apps. A high degree of personalization, ensuring patients trust that the therapy plan is addressing their individual needs, is therefore even more crucial in unguided digital interventions.

On the other hand, as reviewed by Wichers et al. (2011), many people who experience depressive symptoms do not seek treatment, and while this reservation may partly stem from the complexity to seek mental health care and the stigma still associated with depression, evidence suggests that the most common reason depressed people avoid seeking treatment is a desire to manage their problems independently (van Beljouw et al., 2010). Furthermore, in many countries, psychotherapy is still costly and not routinely available or reimbursed by mental health insurance (WHO, 2017; van Beljouw et al., 2010). From this perspective digital treatment applications can provide an inexpensive and direct option for depressed patients to address their problems independently.

A promising strategy for a faster development of personalized psychotherapy may therefore be to create and study the effectiveness of individual therapy plans in digital IMI applications and later transfer these insights into face-to-face psychotherapy. Methodologically, digital applications could use EMA to capture the dynamic relationship between individual target variables and their potential influencing factors, as illustrated by (Harnas et al., 2021), and combine the data derived from such analyses with other decision criteria, such as patients' traits, goals and treatment experiences, in machine learning algorithms that formulate individual adaptive treatment plans.

Enhancing our understanding of depression with EMA

Our findings in Study III show that RNT thoughts, and specifically their level of concreteness, contain information that explains variance in depressive symptoms beyond that explained by Likert-scaled RNT items. Nonetheless, the resource demands of the rating process used in our study, which includes both rater training and the actual rating process, create significant barriers to its use in clinical research and practice. Therefore, future research should emphasize the development of reliable automated approaches for assessing concreteness. In the discussion of Study III, we outlined potential methods for achieving this, including dictionary-based approaches (Pennebaker et al., 2015; Brysbaert et al., 2014) and artificial intelligence (Stade et al., 2024; Lan, Cheng, Sheng, Gao, & Li, 2024; Yang et al., 2023).

Dictionary-based text analysis tools like the ‚Linguistic Inquiry and Word Count' (Pennebaker et al., 2015) quantify the frequency of words in a text, which are pre-assigned to specific psychological, linguistic categories. However, while such rule-based approaches benefit from a high degree of standardisation, the validity of their ratings is limited by the contextual usage of words. For example, generalizing terms such as "always" or "never" are abstract (e.g., "My neighbor never greets me"), but in a context where one is describing another person's statement (e.g., "My neighbor said I would never greet him, but often he doesn't hear my greetings."), it is misleading to rate the concreteness of the sentence by counting its inclusion of abstract versus concrete words.

In contrast, the potential of artificial intelligence is more promising for creating economical and valid concreteness ratings. As the sheer volume of user-generated text on the digital landscape, including social media posts and product reviews, continues to grow beyond the capacity for manual analysis, artificial intelligence, particularly natural language

processing (NLP), has become an essential tool for automatically extracting insights from so called ‚user-generated data' (Kheiri & Karimi, 2023; Furukawa et al., 2023; Burger et al., 2021). In fields such as marketing and politics, NLP techniques are widely used for sentiment analysis, i.e. to evaluate users' emotions and attitudes towards products and topics from the content they generate. Specifically, these analyses rely on machine learning algorithms that use NLP to extract sentiments embedded in textual data (Kheiri & Karimi, 2023).

Recently, also researchers from the field of clinical psychology have begun to explore the potential of NLP for analysing patient-generated text. For instance, Burger et al. (2021) trained an NLP model to identify schemas from underlying thought record, achieving a substantial agreement between the NLP and two manual ratings of Cohen's $\kappa = 0.79$. Another example is the study of Shin et al. (2023), which investigated the potential of large language models (specifically they used ChatGPT from OpenAI; OpenAI, 2024) in detecting depression through user-generated diary text that was journaled by healthy and mildly depressed patients daily within an EMA application. A GPT-3.5 fine tuning model that was trained with a small dataset of about 400 text units achieved an impressive prediction accuracy of about 90%. These results suggest that NLP and particularly large language models can offer a practical, automated solution to generate valid ratings of psychological constructs out of patient-generated text, such as of the concreteness of patients' journaled RNT.

Conclusion

Depression is a highly prevalent (GBD 2019 Mental Disorders Collaborators, 2022), heterogeneous (Fried & Nesse, 2015), and dynamic mental disorder (van Genugten et al., 2022). In addition to long-established clinical assessment techniques like questionnaires and clinical interviews, ecological momentary assessment (EMA) is gaining rapid popularity in mental health research and practice (Colombo, Fernández-Álvarez et al., 2019). With the widespread availability of smartphones across the globe (Statista, 2024), EMA offers the potential to reach many individuals and integrate seamlessly into their daily lives. Its ability to assess symptoms in real-time is of particular value in affective disorders, as the retrospective recall of affect involves several cognitive processes that can bias the memory (Colombo, Fernández-Álvarez et al., 2019; Wichers et al., 2011). Moreover, EMA allows insights into the temporal dynamics between depressive symptoms and underlying processes as they naturally occur (Moberly & Watkins, 2008). This enables new advancements in

psychotherapeutic research and practice, through improving our clinical assessments, treatments, and a deeper understanding of depression.

Findings from EMA studies also support a shift in our perspective on mental health disorders, moving towards a more dynamic and nuanced understanding of these conditions. In cognitive models of depression (Beck, 1979), dynamic relationships between individual symptoms - such as the cyclical interactions between dysfunctional thoughts, negative emotions, and behavioral reactions - have long been postulated. Yet, it is the growing body of EMA research that finally provides ecological empirical evidence for these dynamics. Ultimately, this is driving a shift in our understanding of the nature of mental disorders, including depression (Fried & Cramer, 2017). While our entire categorical diagnostic system originally rests on the assumption that the symptoms of a disorder stem from a common latent brain disease, modern theories suggest that mental disorders are, in fact, networks of mutually influencing factors. These manifestations are not categorical but exist on a continuum (Hofmann et al., 2016). Fried & Cramer (2017) propose a network structure in which some symptoms are interconnected within a network, while others reside in an „external field“. This external field encompasses factors that influence the network system from the outside and may change more slowly - such as metacognitive beliefs, attributional styles, environmental stressors, and neurological factors - which can impact the more rapidly fluctuating symptoms of the network without being directly integrated into the network structure themselves. This distinction underscores the complexity of symptom interactions and emphasizes the importance of considering both state-like symptoms, which are highly dynamic and intercorrelated, and trait-like external factors, which influence the network in a broad manner, when studying depression.

Like most mental disorders, depression shares some of its potential symptoms with other diagnoses (WHO, 2022). Sleep problems for example are the core symptoms of insomnia, changes in appetite are the core symptoms of eating disorders, and concentration problems can also occur in GAD. Erasing the artificial borders between diagnoses, the network perspective explains the frequent comorbidity of depression with other mental disorders and highlights the importance of transdiagnostic factors, such as RNT. In addition, transdiagnostic, global measures of change, such as global functioning and quality of life, are becoming increasingly important for evaluating intervention effects. The network perspective also supports the development of personalized treatments, as symptom profiles are no longer classified into disorders, but observed at an individual level.

From this perspective, it is likely that EMA will play an increasingly important role in clinical research and practice in the future. In the context of psychotherapy, the technical possibilities of EMA, passive data tracking, and artificial intelligence are still in their infancy. Nonetheless, the significant research interest in these technologies suggests that their potential will be rapidly explored, promising disruptive innovations in both research and psychotherapy practice. These rapid advancements, however, underscore the importance of carefully considering and empirically investigating the opportunities and challenges associated with these techniques.

Zusammenfassung

Möglichkeiten und Herausforderungen des Ecological Momentary Assessments für die Psychotherapieforschung bei Depression

Depression – eine hochgradig prävalente und komplexe Störung

Die Depression ist eine hochgradig prävalente und komplexe psychische Störung. Laut den neuesten Schätzungen leben etwa 300 Millionen Menschen, also mehr als 5 % der globalen erwachsenen Bevölkerung, derzeit mit Depressionen (GBD 2019 Mental Disorders Collaborators, 2022; Arias-de la Torre et al., 2021). Trotz der Existenz wirksamer Behandlungen, einschließlich Antidepressiva und Psychotherapie (Cuijpers, Oud et al., 2021; Cuijpers, Miguel et al., 2023), bleiben fast 50% der Patient*innen unbehandelt (Mekonen et al., 2021), und von denjenigen, die leitfadengerechte Behandlungen erhalten, schlagen nur etwa 40% auf die Therapie an (Cuijpers, Karyotaki et al., 2021). Die Mechanismen der Depression besser zu verstehen, die Verteilung der psychischen Gesundheitsversorgung zu verbessern sowie wirksamere Behandlungen zu entwickeln, sind globale Herausforderungen (WHO, 2017).

Die Komplexität der Depression ist gekennzeichnet durch eine hohe Heterogenität der individuellen Symptomprofile (Fried & Nesse, 2015) sowie eine hohe Prävalenz komorbider Störungen, wie z.B. Angststörungen (McGrath et al., 2020). Etwa 40-70 % der depressiven Personen erfüllen gleichzeitig die Kriterien mindestens einer Angststörung (Lamers et al., 2011; Kessler et al., 2015). Eine dritte Komplexitätsstufe liegt auf individueller Ebene, da insbesondere affektive Symptome und kognitive Prozesse der Depression, wie Repetitives Negatives Denken (engl.: repetitive negative thinking, RNT), sowohl zwischen Tagen als auch innhalb von Tagen hochdynamisch sind (van Genugten et al., 2022; Rosenkranz et al., 2020). RNT beschreibt den Prozess des Grübelns oder sich Sorgen-Machens, definiert als negative Gedanken, die repetitiv, intrusiv und schwer kontrollierbar sind (Ehring & Watkins, 2008). RNT ist ein transdiagnostischer Prozess, der eine wichtige Rolle bei der Entstehung und Aufrechterhaltung von Depressionen und Angststörungen spielt (Ehring & Watkins, 2008; Wahl et al., 2019; Egan et al., 2024).

Ecological Momentary Assessment – Depressionen in Echtzeit erforschen

Mit der weitverbreiteten Nutzung von Smartphones (Statista, 2024) gewinnen digitale Techniken zur Erhebung psychischer Symptome wie das Ecological Momentary Assessment (EMA), auch bekannt als Experience Sampling Method (ESM), in der klinischen Forschung und Praxis rasch an Popularität. EMA bezeichnet die hochfrequente Erhebung von Erleben, Verhalten und physiologischen Prozessen in Echtzeit. Diese Technik ist besonders wertvoll für die Erfassung affektiver Symptome, da retrospektive Erhebungen affektiver Erlebnisse

anfällig für Erinnerungsverzerrungen sind, insbesondere bei depressiven Personen (Colombo et al., 2020; Gotlib & Joormann, 2010).

Ziele dieser Dissertation

Diese Arbeit widmet sich den Möglichkeiten von EMA, die Psychotherapieforschung und -praxis bei Depressionen in drei verschiedenen Bereichen zu unterstützen: Die Verbesserung der klinischen Diagnostik, die Beschleunigung der Entwicklung wirksamerer und insbesondere personalisierter psychotherapeutischer Therapien und die Vertiefung unseres Verständnisses von Depressionen.

Um diese Bereiche zu untersuchen, wurde eine EMA-Studie durchgeführt, die in eine große randomisierte kontrollierte Studie eingebettet war, in der die Wirksamkeit von drei psychotherapeutischen Ansätzen zur Behandlung von Depressionen untersucht wurde: Kognitive Verhaltenstherapie (engl.: cognitive behavioral therapy, CBT), Schematherapie (ST) und Individuell-Supportive Therapie (IST; Kopf-Beck et al., 2024). An der EMA Substudie nahmen anfangs N = 106 moderat bis schwer depressive Personen teil, die entweder stationär oder tagklinisch am Max-Planck-Institut für Psychiatrie in München, Deutschland, behandelt wurden. Alle drei Interventionen dauerten sieben Wochen und umfassten zwei Gruppensitzungen (je 100 Minuten) und zwei Einzelsitzungen (je 50 Minuten) pro Woche sowie psychiatrische Standardbehandlungen wie Pharmakotherapie und ergänzende Therapie wie Ergotherapie, für die in den statistischen Analysen kontrolliert wurde. Die EMA-Erhebung wurde zusätzlich zu einer umfassenden Testbatterie durchgeführt, die unter anderem wöchentliche Fragebogenerhebungen (engl.: weekly questionnaire assessments, WQA) zu depressiven Symptomen und RNT sowie klinische Interviews zur globalen Funktionsfähigkeit umfasste (Kopf-Beck et al., 2020). Sie fand dreimal täglich über den gesamten Interventionszeitraum statt und umfasste drei Variablen: ‚momentane depressive Symptome' und ‚momentanes RNT' (Summenscores aus jeweils 4 Likert-skalierten Items) sowie RNT-Gedanken, die über ein Freitext-Item erfasst wurden. Die erhobenen Daten wurden in drei verschiedenen Studien mit unterschiedlichen Forschungsschwerpunkten analysiert:

Studie I

Erste Studien deuten darauf hin, dass EMA die Veränderung von depressiven Symptomen reliabler erfasst und Interventionseffekte in klinischen Studien sensitiver

detektiert. Allerdings mangelt es an der empirischen Untersuchung dieser Annahmen (Moore et al., 2016) sowie der Vergleichbarkeitkeit von EMA und Fragebögen hinsichtlich ihrer Vorhersage von klinischen Interview-Ergebnissen (Targum et al., 2021).

Daher haben wir in Studie I die Vergleichbarkeit zwischen EMA und WQA in ihrer Erfassung der Veränderung depressiver Symptome und RNT untersucht. Wir analysierten: a) die Größe der mit beiden Techniken verbundenen Interventionseffekte, und b) ihre Validität bezüglich der Vorhersage von klinischen Interview-Ergebnissen zur Veränderung des globales Funktionsniveaus.

Im Einklang mit den Ergebnissen von Moore et al. (2016) fanden wir, dass EMA signifikante Interventionseffekte zwischen den Interventionsgruppen erkannte, die durch WQA nicht identifiziert wurden. Konkret offenbarte EMA, dass ST RNT effektiver reduzierte als die beiden anderen Interventionsgruppen, CBT und IST. Im Kontrast dazu, sagte WQA die Veränderungen des globalen Funktionsniveaus präziser vorher, obwohl auch die mit EMA erfasste Veränderung der depressiven Symptome signifikante Vorhersagen lieferte. Zudem stellten wir signifikante Unterschiede in den Zeiteffekten (Slopes) zwischen den beiden Erhebungstechniken fest: WQA zeigte eine steilere Reduktion der depressiven Symptome und RNT über die Zeit und extremere Werte zu Beginn und Ende der Therapie, d.h. höhere Anfangs- und niedrigere End-Werte im Vergleich zu EMA.

Diese Ergebnisse legen nahe, dass EMA’s höhere Sensitivität für die Detektierung von Therapieffekten zwischen Interventionsbedingungen nicht auf eine höhere Sensitivität für Veränderungen zurückzuführen ist. Stattdessen gehen wir davon aus, dass EMA Veränderungen depressiver Symptome reliabler erfasst, indem Erinnerungsverzerrungen vermieden werden (Colombo, Suso-Ribera et al., 2019), für die retrospektive Fragebögen anfällig sind. Dies ermöglicht es EMA kleine Unterschiede zwischen Interventionseffekten mit höherer statistischer Power zu identifizieren. Darüber hinaus deuten unsere Ergebnisse darauf hin, dass retrospektive Fragebögen möglicherweise die Amplitude zwischen der Symptomschwere zu Beginn und Ende der Behandlung, d.h. den Interventionseffekt, überschätzen. Erinnerungsverzerrungen, kommen zwar bei allen Menschen vor, treten bei Personen mit Depressionen jedoch besonders stark auf, insbesondere bezogen auf die Überschätzung negativer Affekte (Colombo, Suso-Ribera et al., 2019; Colombo et al., 2020). Depressive Personen könnten ihre Symptome daher insbesondere zu Beginn der Therapie, wenn ihre Depression noch vergleichsweise schwer ist, im Vergleich zum Ende der Therapie,

wenn sich ihre Symptome durch die Therapie verbessert haben, überschätzen.

Für die Interpretation dieser Ergebnisse ist es wichtig, in Folgestudien systematisch zu untersuchen, wie sich Erinnerungsverzerrungen im Therapieverlauf verändern (Ebner-Priemer & Trull, 2009b). Eine weitere Erklärung für die gefundenen Ergebnisse könnte das EMA-Design unserer Studie sein. Die hochfrequenten repetitiven Messungen könnten ebenfalls Verzerrungen eingeführt haben, wie z.B. unzuverlässige monotone Antworten aufgrund einer Ermüdung der Teilnehmer. Dass EMA durch mit klinischen Interviews erhobene signifikante Veränderungen im globalen Funktionsniveau vorhersagen kann, unterstützt die externe Validität von EMA. Eine Erklärung für die überlegene Vorhersagegenauigkeit der Fragebögen könnte die gemeinsame Retrospektivität von klinischen Interviews und Fragebögen sein.

Studie II

Da sich EMA vergleichsweise einfach in den Alltag von Personen integrieren lässt und die frequente Messung von Veränderungen ermöglicht (Colombo, Fernández-Álvarez et al., 2019; Trull & Ebner-Priemer, 2009b), ist es eine vielversprechende Technik zur Überwachung früher Behandlungsfortschritte (Li et al., 2023). In der Pharmakotherapie gibt es klare Richtlinien zur Anpassung der Medikation, wenn in den ersten Behanldungswochen keine ausreichende Besserung beobachtet wird (Gautam et al., 2017). Für die Psychotherapie fehlen solche Richtlinien. Frühe Indikatoren für das Nicht-Ansprechen von Patient*innen auf eine begonnene Behandlung können genutzt werden, um Interventionen nach dem Stepped-Care-Ansatzes (van Straten et al., 2015) anzupassen und begrenzte therapeutische Ressourcen effizienter zu verteilen (Richards, 2012) sowie den Evaluationsprozess klinischer Studien zu personalisierten Therapien zu beschleunigen (Kidwell & Almirall, 2023). Trotz starker Evidenz, die frühe Verbesserungen als robusten Prädiktor für die individuelle Wirksamkeit von Psychotherapie unterstützt (Beard & Delgadillo, 2019; Li et al., 2023), gibt es bislang keinen Konsens darüber, welcher Zeitpunkt und welche Verbesserungsrate der beste Prädiktor für die Vorhersage des Behandlungsergebnisses ist (Beard & Delgadillo, 2019).

In Studie II untersuchten wir daher folgende Forschungsfragen: (1a) Zu welchem Zeitpunkt nach Behandlungsbeginn sagt eine frühe Verbesserung der depressiven Symptome signifikant den Behandlungserfolg voraus? (1b) Sagen sowohl WQA als auch EMA zu diesen Zeitpunkten den Behandlungserfolg voraus? und (2) Wie prädiktiv sind verschiedene Definitionen der frühen Verbesserung in Bezug auf das definierte Zeitfenster und den

Symptom-Cutoff? Zur Untersuchung unserer Forschungsfragen führten wir Lineare Regressionsmodelle und Receiver-Operating-Characteristic-Analysen durch, um den Behandlungserfolg (BDI-II-Verbesserung von vor- bis nach der Intervention um ≥ 50%) vorherzusagen. Darüber hinaus berechneten wir für folgende Definitionen der frühen Verbesserung das Verhältnis zwischen wahr-negativ zu falsch-negativ Vorhersagen, um ihren prädiktiven Wert zu bewerten: 10%, 20%, 30% oder 40% Verbesserung nach einer, zwei, drei oder vier Wochen der Behandlung.

Unsere Ergebnisse in Studie II zeigen, dass sowohl EMA- als auch WQA-gemessene frühe Verbesserungen das Ansprechen (versus Nicht-Ansprechen) auf die sieben-wöchige Behandlung signifikant vorhersagen. Bereits nach drei Wochen Behandlung lieferten beide Erhebungsmethoden signifikante Vorhersagen mit vergleichbaren Area-under-the-Curve (AUC)-Werten von 73% (EMA) und 77% (WQA). Der Fragebogenprädiktor zeigte jedoch ein klareres Muster in bezug auf die Veränderungsrate, die als bester Prädiktor für klinische Implikationen dienen könnte - in unserer Studie definiert als ein hohes Verhältnis zwischen der richtig-negativ und der falsch-negativ Rate. Der beste Prädiktor war eine durch WQA gemessene 10%ige Verbesserung nach vier Behandlungswochen, was zu einer wahr-negativ Rate von 22% im Vergleich zu einer falsch-negativ Rate von 0% führte.

Es ist wichtig zu beachten, dass der Fragebogen-Prädiktor sowie die abhängige Variable ‚Behandlungserfolg' mit demselben Fragebogen (BDI-II) operationalisiert wurden. Da diese Übereinstimmung dem Fragebogen-Prädiktor möglicherweise einen Vorteil verschafft, ist es überraschend, dass die EMA-Vorhersagen ähnlich präzise waren. Darüber hinaus könnte dieser potenziell „unfaire" Vergleich erklären, warum das Vorhersagemuster des Fragebogens bezüglich der prädiktivsten Veränderungsrate klarer war als das von EMA. Zukünftige Studien sollten daher eine „neutralere" abhängige Variable verwenden, die mit keinem der beiden Prädiktoren überlappt, wie z.B. die globale Funktionsfähigkeit oder Lebensqualität.

Studie III

Studie III untersuchte mit EMA die zeitlichen Zusammenhänge zwischen momentanen depressiven Symptomen und RNT (Ehring & Watkins, 2008; Rosenkranz et al., 2020) sowie deren Veränderungen im Therapieverlauf. Anders als in den meisten EMA-Studien zu RNT, konzentrierten wir uns speziell auf die Konkretheit (Stöber & Borkovec, 2002) von RNT, indem die Patienten dreimal täglich ihre RNT-Gedanken aufschrieben, deren

Konkretheitslevel anschließend von geschulten Beurteilern bewertet wurden. Studien zeigen, dass die Grübel- und Sorgen-Gedanken depressiver Personen im Mittel abstrakter, als die von gesunden Personen sind, d.h. unklarer, verallgemeinernder, situationsübergreifender und weniger lösungsorientiert (Joormann et al., 2006; Stöber & Borkovec, 2002). Während gut belegt ist, dass psychotherapeutische Ansätze wie CBT RNT wirksam reduzieren (Bell et al., 2023), wurden Veränderungen in der Konkretheit von RNT im Verlauf der Psychotherapie bislang noch nicht auf einer momentanen Ebene untersucht. Zudem wurden die zeitlichen Dynamiken momentaner depressiver Symptome bislang nur im Zusammenhang mit momentanem RNT, nicht jedoch mit der Konkretheit von momentanem RNT untersucht (Watkins & Moulds, 2005a).

Daher untersuchten wir in Studie III folgende Forschungsfragen: (1) Erklärt die Konkretheit als spezifischer Modus des momentanen RNT zusätzliche Varianz in der Vorhersage momentaner depressiver Symptome, die über die durch momentanes RNT an sich aufgeklärte Varianz hinausgeht? (2) Nimmt die Konkretheit des momentanen RNT im Verlauf der Psychotherapie zu? und (3) Wie hängen momentane depressive Symptome mit der Konkretheit von momentanem RNT zeitlich zusammen? Geht ein Faktor dem anderen zeitlich voraus? Die Hypothesen wurden mittels Multi-Level Modellen (MLM) getestet.

Studie III zeigt, dass ein Modell, das sowohl RNT als auch die Konkretheit berücksichtigt, signifikant mehr Varianz in momentanen depressiven Symptomen erklärt als ein Modell, dass nur RNT berücksichtigt. Überraschenderweise nahm die Konkretheit von RNT im Verlauf der Therapie insgesamt ab. Genauer interagierte die Veränderung der RNT-Konkretheit mit der Veränderung der Depressionsschwere: Bei Personen, deren Depression sich überdurchschnittlich verbesserte, nahm die RNT-Konkretheit leicht zu, während sie bei Personen, die sich durchschnittlich oder unterdurchschnittlich verbesserten, abnahm. Außerdem zeigte sich, dass höhere Werte momentaner depressiver Symptome nachfolgende Reduktionen der RNT-Konkretheit vorhersagten, jedoch nicht umgekehrt, und dass diese Dynamik über den Verlauf der Therapie stabil blieb.

Die Ergebnisse aus Studie III zeigen, dass die momentanen depressiven Symptome der Personen in einzigartiger Weise sowohl mit dem RNT-Prozess als auch mit der Konkretheit des RNT-Inhalts verbunden sind. Um valide Schlussfolgerungen aus dem Befund zu ziehen, dass sich die RNT-Konkretheit im Verlauf der Behandlung in Abhängigkeit vom Schweregrad der Depression verändert, sollten zukünftige Studien Mediationsanalysen (Baron & Kenny, 1986) durchführen, um zu untersuchen, ob die momentane RNT-

Konkretheit als Mechanismus der Veränderung in psychologischen Behandlungen von Depression dient. Die zeitliche Dynamik zwischen der momentanen RNT-Konkretheit und depressiven Symptomen ist überraschend, da frühere Studien wechselseitige Effekte für momentanes RNT fanden, d.h., höhere negative Affekte sagten spätere RNT-Zunahmen voraus und umgekehrt (Moberly & Watkins, 2008). Der Effekt von Konkretheit auf Depression könnte jedoch verzögert auftreten, d.h. über längere Zeiträume wie Tage oder Wochen, oder sich in bestimmten Momenten zeigen, etwa wenn konkretes Denken zur Lösung eines Problems führt (Watkins & Moulds, 2005a; Watkins, 2008).

Methodische Grenzen und Zukunftsrichtungen zur Verbesserung von EMA Protokollen

Folgende Strategien könnten zu einer weiteren Verbesserung der Datenqualität sowie der Patient*innen Adhärenz in EMA-Studien beitragen: a) Entwicklung dynamischer EMA-Zeitpläne durch variablenspezifische Untersuchung der Fluktuationsrate, Veränderungsgeschwindigkeit und des geeigneten Erhebungs-Modus (momentane oder aggregierte Erfassung) sowie Einbeziehung von ‚planned missing data designs' (Silvia et al., 2013), b) Verwendung validierter EMA-Items (Trull & Ebner-Priemer, 2020), c) passives Messen von Verhaltensvariablen und psychophysiologischen Prozessen (Ebner-Priemer & Trull, 2009b), d) neutrale Itemformulierungen und Ermöglichung bildschirmfreier Zeiten (Heo et al., 2017), um negative reaktive EMA-Effekte zu vermeiden, und e) die Untersuchung, ob und unter welchen Bedingungen reaktive EMA-Effekte auftreten (Domhardt et al., 2021; Ballegooijen et al., 2016).

Weitere klinische Implikationen und Zukunftsrichtungen für die Integration von EMA in die Psychotherapie-Forschung bei Depression

Sowohl retrospektive Fragebögen als auch EMA und klinische Interviews haben einzigartige Stärken und Schwächen. Die Kombination aller drei Erhebungsverfahren könnte Kliniker*nnen helfen, die depressiven Symptome ihrer Patient*innen besser zu verstehen sowie präzisere Ergebnisse in klinischen Studien ermöglichen. Klinische Interviews könnten sich am besten für die Diagnosestellung von Depressionen eigenen, EMA für die Überwachung hochdynamischer depressiver Symptome und Fragebögen zur Erfassung stabiler Eigenschaften, die zuverlässig retrospektiv berichtet werden können.

Darüber hinaus kann EMA über zwei Hauptwege zur Entwicklung effektiverer und insbesondere personalisierter Psychotherapien beitragen: a) durch die Beschleunigung des

Evaluationsprozesses klinischer Studien, wie durch die Strategien in Studie I und II dargestellt, und b) durch die Verbesserung der Wirksamkeit von Therapien, z.B. durch die Bereitstellung personalisierten Feedbacks und Echtzeit-Empfehlungen (Wichers et al., 2011; Colombo, Fernández-Álvarez et al., 2019) sowie durch die Zuweisung von Therapiemodulen auf Basis kontinuierlicher EMA-Daten (Harnas et al., 2021).

In der Marktforschung werden häufig sogenannte Sentiment-Analysen durchgeführt (Kheiri & Karimi, 2023), die automatisiert nutzergenerierte Inhalte untersuchen, um Informationen über die Einstellungen von Nutzern zu Produkten oder Themen zu gewinnen. Studie III zeigt, dass die Analyse von nutzergenerierten Inhalten, in diesem Fall die notierten Gedanken der Patient*innen, auch in der Psychotherapieforschung Informationen liefert, die über den Informationsgehalt Likert-skalierter Items hinaus gehen. Da die manuellen Konkretheitsbewertungen, die in unserer Studie von geschulten externen Bewertern durchgeführt wurden, sehr zeit- und ressourcenintensiv sind, sollten zukünftige Studien versuchen, auf natürlichen Sprachmodellen (engl.: natural language processing; NLP) basierende Programme zu entwickeln, die eine automatisierte Bewertung der Konkretheit von RNT sowie anderer mit Depression assoziierter Gedankenmerkmale ermöglichen (Shin et al., 2023).

Fazit

Die Möglichkeit mit EMA, Symptome in Echtzeit zu erfassen, ist besonders wertvoll bei affektiven Störungen, bei denen die retrospektive Erinnerung verzerrt sein kann (Colombo, Fernández-Álvarez et al., 2019; Wichers et al., 2011). Darüber hinaus zeigen EMA-Studien Symptomdynamiken auf, die in kognitiven Modellen der Depression schon lange postuliert werden (Beck, 1979). Sie unterstützen damit auch einen Wandel unseres Verständnisses von Depressionen, weg von der Annahme, dass die Symptome eines Störungsbilds einer latenten Krankheit entspringen, hin zu einer Netzwerkperspektive (Fried and Cramer's, 2017). Eine Netzwerkperspektive klärt die hohe Prävalenz von Komorbiditäten, hebt die Bedeutung transdiagnostischer Faktoren wie RNT und die globale Funktionsfähigkeit hervor und unterstützt personalisierte Therapieansätze. Angesichts der rasanten Verbreitung neuer Technologien zur Erfassung und Behandlung depressiver Symptome, wie EMA, passivem Daten-Tracking und Künstlicher Intelligenz, ist es jedoch wichtig, die Möglichkeiten aber auch Herausforderungen dieser Technologien zur Erfassung psychologischer Variablen kontinuierlich und sorgfältig zu untersuchen.

Abbreviations

AUC	Area Under the Curve
CBT	Cognitive Behavioral Therapy
EMA	Ecological Momentary Assessment
EMI	Ecological Momentary Intervention
GAD	General Anxiety Disorder
IMI	Internet- and Mobile-based Interventions
IST	Individual Supportive Therapy
HRV	Heart Rate Variability
LRM	Linear Regression Model
MLM	Multilevel Model
NNT	Number Needed to Treat
OPTIMA	Optimized Psychotherapy Identification at the Max-Planck-Institute of Psychiatry
RCT	Randomized Controlled Trial
rfCBT	Rumination-focused cognitive behavioral therapy
ROC	Receiver Operating Characteristic
RR	Responder rate
ST	Schema Therapy
TNFN	Ratio between the true negative rate and the false negative rate
wTNFN	Ratio between the true negative rate and the false negative rate multiplied (weighted) by the ratio between the non-responder rate and the responder rate
WQA	Weekly Questionnaire Assessment

References

Altan-Atalay, A., Kaya-Kızılöz, B., İlkmen, Y. S., & Kozol, E. (2022). Impact of abstract vs. Concrete processing on state rumination: An exploration of the role of cognitive flexibility. *Journal of Behavior Therapy and Experimental Psychiatry*, 74, 101691. https://doi.org/10.1016/j.jbtep.2021.101691.

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, *63*(2020), 10. https://doi.org/10.1007/b97636.

Angel, V. de, Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D. C., Dobson, R., & Hotopf, M. (2022). Digital health tools for the passive monitoring of depression: a systematic review of methods. *Npj Digital Medicine*, *5*(1), 3. https://doi.org/10.1038/s41746-021-00548-8.

Arditte, K. A., Shaw, A. M., & Timpano, K. R. (2016). Repetitive Negative Thinking: A Transdiagnostic Correlate of Affective Disorders. *Journal of Social and Clinical Psychology*, 35(3), 181–201. https://doi.org/10.1521/jscp.2016.35.3.181.

Arias-de la Torre, J., Vilagut, G., Ronaldson, A., Serrano-Blanco, A., Martín, V., Peters, M., ... & Alonso, J. (2021). Prevalence and variability of current depressive disorder in 27 European countries: a population-based study. *The Lancet Public Health*, 6(10), e729-e738. https://doi.org/10.1016/S2468-2667(21)00047-5.

Arias, D., Saxena, S., & Verguet, S. (2022). Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine*, 54. https://doi.org/10.1016/j.eclinm.2022.101675.

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D. D., Smit, J. H., & Riper, H. (2016). Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): protocol of a randomised controlled trial. *BMC psychiatry*, *16*, 1-6. https://doi.org/10.1186/s12888-016-1065-5.

Barbato, A., Vallarino, M., Rapisarda, F., Lora, A., & de Almeida, J. M. C. (2016). EU compass for action on mental health and well-being. Access to mental health care in Europe. Scientific paper. Funded by the European Union in the frame of the 3rd EU Health Programme (2014–2020).

Barge-Schaapveld, D. Q. C. M., & Nicolson, N. A. (2002). Effects of antidepressant treatment on the quality of daily life: an experience sampling study. *Journal of Clinical Psychiatry*, *63*(6), 477-485. https://doi.org/10.4088/jcp.v63n0603.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173. https://doi.org/10.1037/0022-3514.51.6.1173.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. (67(1)), 1–48. https://doi.org/10.18637/jss.v067.i01.

Beck, A. T. (1979). Cognitive therapy of depression. Guilford press.

Beck, A. T. (2002). Cognitive models of depression. *Clinical Advances in Cognitive Psychotherapy: Theory and Application*, *14*(1), 29–61.

Beck, J. S. (2021). Cognitive behavior therapy: Basics and beyond (Third edition). New York: The Guilford Press.

Behar, E., McGowan, S. K., McLaughlin, K. A., Borkovec, T. D., Goldwin, M., & Bjorkquist, O. (2012). Concreteness of positive, negative, and neutral repetitive thinking about the future. *Behavior Therapy*, 43(2), 300–312. https://doi.org/10.1016/j.beth.2011.07.00.

Bell, I. H., Marx, W., Nguyen, K., Grace, S., Gleeson, J., & Alvarez-Jimenez, M. (2023). The effect of psychological treatment on repetitive negative thinking in youth depression and anxiety: A meta-analysis and meta-regression. *Psychological Medicine*, 53(1), 6-16. https://doi.org/10.1017/S0033291722003373.

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion*, *23*(5), 1021–1040. https://doi.org/10.1080/02699930802607937.

Beard, J. I. L., & Delgadillo, J. (2019). Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: A systematic review and meta-analysis. *Depression and Anxiety*, 36(9), 866–878. https://doi.org/10.1002/da.22931.

Blackwell, S. E., Woud, M. L., Margraf, J., & Schönbrodt, F. D. (2019). Introducing the leapfrog design: A simple Bayesian adaptive rolling trial design for accelerated treatment development and optimization. *Clinical Psychological Science*, 7(6), 1222-1243. https://doi.org/10.1177/2167702619858071.

Blanke, E. S., Neubauer, A. B., Houben, M., Erbas, Y., & Brose, A. (2022). Why do my thoughts feel so bad? Getting at the reciprocal effects of rumination and negative affect using dynamic structural equation modeling. *Emotion*, *22*(8), 1773. https://doi.org/10.1037/emo0000946.

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological bulletin*, *148*(7-8), 588. https://doi.org/10.1037/bul0000365.

Bolin, E., & Lam, W. (2013). A review of sensitivity, specificity, and likelihood ratios: evaluating the utility of the electrocardiogram as a screening tool in hypertrophic cardiomyopathy. *Congenital heart disease*, 8(5), 406-410. https://doi.org/10.1111/chd.12083.

Bondolfi, G., Jermann, F., Rouget, B. W., Gex-Fabry, M., McQuillan, A., Dupont-Willemin, A., ... & Nguyen, C. (2010). Self-and clinician-rated Montgomery–Åsberg Depression Rating Scale: evaluation in clinical practice. *Journal of affective disorders*, *121*(3), 268-272. https://doi.org/10.1016/j.jad.2009.06.037.

Borkovec, T. D., Robinson, E., Pruzinsky, T., & DePree, J. A. (1983). Preliminary exploration of worry: Some characteristics and processes. *Behaviour Research and Therapy*, 21(1), 9–16. https://doi.org/10.1016/0005-7967(83)90121-3

Borkovec, T. D., Ray, W. J., & Stöber, J. (1998). Worry: A cognitive phenomenon intimately linked to affective, physiological, and interpersonal behavioral processes. *Cognitive Therapy and Research*, 22, 561–576. https://doi.org/10.1023/A:1013845821848.

Borkovec, T. D., Shadick, R. N., & Hopkins, M. (1991). The nature of normal and pathological worry. Retrieved from https://psycnet.apa.org/record/1991-98548-002.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5.

Buckman, J. E., Underwood, A., Clarke, K., Saunders, R., Hollon, S. D., Fearon, P., & Pilling, S. (2018). Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis. *Clinical psychology review*, *64*, 13-38. https://doi.org/10.1016/j.cpr.2018.07.005.

References

Bundesärztekammer, Kassenärztliche Bundesvereinigung & Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (2022). *Nationale VersorgungsLeitlinie Unipolare Depression - Langfassung, Version 3.2*. https://doi.org/10.6101/AZQ/000505.

Burger, F., Neerincx, M. A., & Brinkman, W. P. (2021). Natural language processing for cognitive therapy: extracting schemas from thought records. *PloS one*, *16*(10), e0257832. https://doi.org/10.1371/journal.pone.0257832.

Busch, P. A., & McCarthy, S. (2021). Antecedents and consequences of problematic smartphone use: A systematic literature review of an emerging research area. *Computers in human behavior*, *114*, 106414. https://doi.org/10.1016/j.chb.2020.106414.

Chen, Z., Zhao, S., Tian, S., Yan, R., Wang, H., Wang, X., Zhu, R., Xia, Y., Yao, Z., & Lu, Q. (2022). Diurnal mood variation symptoms in major depressive disorder associated with evening chronotype: Evidence from a neuroimaging study. *Journal of Affective Disorders*, *298*, 151–159. https://doi.org/10.1016/j.jad.2021.10.087

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences and eii*. Hillsdale NJ Eribaum. https://doi.org/10.4324/9780203771587

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology*, *14*(1), 209-236. https://doi.org/10.1146/annurev-clinpsy-050817-084746.

Colombo, D., Suso-Ribera, C., Fernandez-Álvarez, J., Felipe, I. F., Cipresso, P., Palacios, A. G., & Riva, Giuseppe & Botella, Cristina (Eds.) (2019). *Exploring affect recall bias and the impact of mild depressive symptoms: an ecological momentary study*. Springer. https://doi.org/10.1007/978-3-030-25872-6_17

Colombo, D., Suso-Ribera, C., Fernández-Álvarez, J., Cipresso, P., Garcia-Palacios, A., Riva, G., & Botella, C. (2020). Affect Recall Bias: Being Resilient by Distorting Reality. *Cognitive Therapy and Research*, *44*(5), 906–918. https://doi.org/10.1007/s10608-020-10122-3

Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., García-Palacios, A., Cipresso, P., Riva, G., & 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. https://doi.org/10.3390/jcm8040465

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A Procedure for Evaluating Sensitivity to Within-Person Change: Can Mood Measures in Diary Studies Detect Change Reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

Crits-Christoph, P., Connolly, M. B., Gallop, R., Barber, J. P., Tu, X., Gladis, M., & Siqueland, L. (2001). Early improvement during manual-guided cognitive and dynamic psychotherapies predicts 16-week remission status. *The Journal of Psychotherapy Practice and Research*, 10(3), 145. https://pubmed.ncbi.nlm.nih.gov/11402077/.

Cuijpers, P., Huibers, M., Ebert, D. D., Koole, S. L., & Andersson, G. (2013). How much psychotherapy is needed to treat depression? A metaregression analysis. *Journal of affective disorders*, 149(1-3), 1-13. https://doi.org/10.1016/j.jad.2013.02.030

Cuijpers, P., Ebert, D. D., Acarturk, C., Andersson, G., & Cristea, I. A. (2016). Personalized psychotherapy for adult depression: a meta-analytic review. *Behavior Therapy*, *47*(6), 966-980. https://doi.org/10.1016/j.beth.2016.04.007.

Cuijpers, P. (2015). Psychotherapies for adult depression: recent developments. *Current Opinion in Psychiatry*, 28(1), 24–29. https://doi.org/10.1097/YCO.0000000000000121

Cuijpers, P., Karyotaki, E., Ciharova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021). The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta-analysis. *Acta Psychiatrica Scandinavica*, 144(3), 288–299. https://doi.org/10.1111/acps.13335

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Ebert, D., & Karyotaki, E. (2023). Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: A comprehensive meta-analysis including 409 trials with 52,702 patients. *World Psychiatry*, *22*(1), 105-115. https://doi.org/10.1002/wps.21069.

References

Cuijpers, P., Oud, M., Karyotaki, E., Noma, H., Quero, S., Cipriani, A., ... & Furukawa, T. A. (2021). Psychologic treatment of depression compared with pharmacotherapy and combined treatment in primary care: a network meta-analysis. *The Annals of Family Medicine*, 19(3), 262-270. https://doi.org/10.1370/afm.2676.

Cuijpers, P., Quero, S., Noma, H., Ciharova, M., Miguel, C., Karyotaki, E., Cipriani, A., Cristea, I. A., & Furukawa, T. A. (2021). Psychotherapies for depression: A network meta-analysis covering efficacy, acceptability and long-term outcomes of all main treatment types. *World Psychiatry*, *20*(2), 283–293. https://doi.org/10.1002/wps.20860

David, D., Cristea, I., & Hofmann, S. G. (2018). Why Cognitive Behavioral Therapy Is the Current Gold Standard of Psychotherapy. *Frontiers in Psychiatry*, 9, 4. https://doi.org/10.3389/fpsyt.2018.00004.

Dell'Acqua, C., Dal Bò, E., Benvenuti, S. M., & Palomba, D. (2020). Reduced heart rate variability is associated with vulnerability to depression. *Journal of Affective Disorders Reports*, *1*, 100006. https://doi.org/10.1016/j.jadr.2020.100006.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845. https://doi.org/10.2307/2531595

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLOS ONE*, 9(1), e83875. https://doi.org/10.1371/journal.pone.0083875.

Domhardt, M., Cuijpers, P., Ebert, D. D., & Baumeister, H. (2021). More light? Opportunities and pitfalls in digitalized psychotherapy process research. *Frontiers in Psychology*, *12*, 544129. https://doi.org/10.3389/fpsyg.2021.544129.

Ebner-Priemer, U. W., & Trull, T. J. (2009a). Ambulatory assessment: An innovative and promising approach for clinical psychology. *European Psychologist*, *14*(2), 109–119. https://doi.org/10.1027/1016-9040.14.2.109.

Ebner-Priemer, U. W., & Trull, T. J. (2009b). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological assessment*, *21*(4), 463. https://doi.org/10.1037/a0017075.

Edge, D., Watkins, E. R., Newbold, A., Ehring, T., Frost, M., & Rosenkranz, T. (2024). Evaluating the Effects of a Self-Help Mobile Phone App on Worry and Rumination Experienced by Young Adults: Randomized Controlled Trial. *JMIR MHealth and UHealth*, 12, e51932. https://doi.org/10.2196/51932.

Egan, S. J., Greene, D., Callaghan, T., Raghav, S., Funk, J., Badenbach, T., ... & Kopf-Beck, J. (2024). Worry and rumination as a transdiagnostic target in young people: a co-produced systematic review and meta-analysis. *Cognitive Behaviour Therapy*, 1-24. https://doi.org/10.1080/16506073.2024.2369936.

Ehring, T., & Watkins, E. R. (2008). Repetitive negative thinking as a transdiagnostic process. *International Journal of Cognitive Therapy, 1*(3), 192–205. https://doi.org/10.1680/ijct.2008.1.3.192

Ehring, T., Zetsche, U., Weidacker, K., Wahl, K., Schönfeld, S., & Ehlers, A. (2011). The Perseverative Thinking Questionnaire (PTQ): Validation of a content-independent measure of repetitive negative thinking. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 225–232. https://doi.org/10.1016/j.jbtep.2010.12.003.

Elhai, J. D., Dvorak, R. D., Levine, J. C., & Hall, B. J. (2017). Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology. *Journal of affective disorders*, 207, 251-259. https://doi.org/10.1016/j.jad.2016.08.030.

Everaert, J., Podina, I. R., & Koster, E. H. W. (2017). A comprehensive meta-analysis of interpretation biases in depression. *Clinical Psychology Review*, 58, 33–48. https://doi.org/10.1016/j.cpr.2017.09.005.

Fava, M., Rush, A. J., Alpert, J. E., Balasubramani, G. K., Wisniewski, S. R., Carmin, C. N., ... & Trivedi, M. H. (2008). Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR* D report. *American Journal of Psychiatry*, *165*(3), 342-351. https://doi.org/10.1176/appi.ajp.2007.06111868.

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soyster, P. D., Diamond, A. E., & Barkin, J. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour research and therapy*, 116, 69-79. https://doi.org/10.1016/j.brat.2019.01.010.

References

Flett, A. L., Haghbin, M., & Pychyl, T. A. (2016). Procrastination and depression from a cognitive perspective: An exploration of the associations among procrastinatory automatic thoughts, rumination, and mindfulness. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *34*, 169-186. https://doi.org/10.1007/s10942-016-0235-1.

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, 99(1), 20. https://doi.org/10.1037/0033-2909.99.1.20

Folkersma, W., Veerman, V., Ornée, D. A., Oldehinkel, A. J., Alma, M. A., & Bastiaansen, J. A. (2021). Patients' experience of an ecological momentary intervention involving self-monitoring and personalized feedback for depression. *Internet Interventions*, *26*, 100436. https://doi.org/10.1016/j.invent.2021.100436.

Foulkes, L., & Andrews, J. L. (2023). Are mental health awareness efforts contributing to the rise in reported mental health problems? A call to test the prevalence inflation hypothesis. *New Ideas in Psychology*, *69*, 101010. https://doi.org/10.1016/j.newideapsych.2023.101010.

Frank, J. D. (1971). Therapeutic factors in psychotherapy. *American Journal of Psychotherapy*, *25*(3), 350–361. https://doi.org/10.1176/appi.psychotherapy.1971.25.3.350

Fried, E. I. (2017). Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert review of neurotherapeutics*, *17*(5), 423-425. https://doi.org/10.1080/14737175.2017.1307737.

Fried, E. I., & Cramer, A. O. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999-1020. https://doi.org/10.1177/1745691617705892.

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR* D study. *Journal of affective disorders*, *172*, 96-102. https://doi.org/10.1016/j.jad.2014.10.010.

Friederich, H. C., Kruse, P., Kruse, J., Kampling, H., Werner, S., Zara, S., ... & Szecsenyi, J. (2024). Outpatient Psychotherapy in Germany: An evaluation of the structural reform.

Deutsches Ärzteblatt International, 121(10). https://doi.org/10.3238/arztebl.m2024.0039.

Friedl, N., Berger, T., Krieger, T., Caspar, F., & Grosse Holtforth, M. (2020). Using the Personalized Advantage Index for individual treatment allocation to cognitive behavioral therapy (CBT) or a CBT with integrated exposure and emotion-focused elements (CBT-EE). *Psychotherapy Research*, *30*(6), 763-775. https://doi.org/10.1080/10503307.2019.1664782.

Funk, J., Kopf-Beck, J., Watkins, E., & Ehring, T. (2023). Does an app designed to reduce repetitive negative thinking decrease depression and anxiety in young people? (RETHINK): A randomized controlled prevention trial. *Trials*, 24(1), 295. https://doi.org/10.1186/s13063-023-07295-z.

Funk, J., Takano, K., Babl, M., Goldstein, R., Oberwestersberger, R., Kopf-Beck, J., . . . Ehring, T. (2024). Can an intervention designed to reduce repetitive negative thinking alter the response to a psychosocial stressor? A randomized controlled study. *Behaviour Research and Therapy*, 178, 104547. https://doi.org/10.1016/j.brat.2024.104547.

Funk, J., Takano, K., Schumm, H., & Ehring, T. (2022). The Bi-factor model of repetitive negative thinking: Common vs. unique factors as predictors of depression and anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, 77, 101781. https://doi.org/10.1016/j.jbtep.2022.101781

Furukawa, T. A., & Leucht, S. (2011). How to obtain NNT from Cohen's d: Comparison of two methods. *PLOS ONE*, *6*(4), e19070. https://doi.org/10.1371/journal.pone.0019070.

Furukawa, T. A., Iwata, S., Horikoshi, M., Sakata, M., Toyomoto, R., Luo, Y., ... & Aramaki, E. (2023). Harnessing AI to Optimize Thought Records and Facilitate Cognitive Restructuring in Smartphone CBT: An Exploratory Study. *Cognitive Therapy and Research*, *47*(6), 887-893. https://doi.org/10.1007/s10608-023-10411-7.

Garratt, G., Ingram, R. E., Rand, K. L., & Sawalani, G. (2007). Cognitive processes in cognitive therapy: Evaluation of the mechanisms of change in the treatment of depression. *Clinical Psychology: Science and Practice*, 14(3), 224. https://doi.org/10.1111/j.1468-2850.2007.00081.x.

References

Gautam, S., Jain, A., Gautam, M., Vahia, V. N., & Grover, S. (2017). Clinical Practice Guidelines for the management of Depression. *Indian Journal of Psychiatry*, 59(Suppl 1), S34-S50. https://doi.org/10.4103/0019-5545.196973.

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, 9(2), 137-150. https://doi.org/10.1016/S2215-0366(21)00395-3.

Ghio, L., Gotelli, S., Marcenaro, M., Amore, M., & Natta, W. (2014). Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. *Journal of affective disorders*, *152*, 45-51. https://doi.org/10.1016/j.jad.2013.10.002.

Gloster, A. T., Rinner, M. T.B., Ioannou, M., Villanueva, J., Block, V. J., Ferrari, G., Benoy, C., Bader, K., & Karekla, M. (2020). Treating treatment non-responders: A meta-analysis of randomized controlled psychotherapy trials. *Clinical Psychology Review*, 75, 101810. https://doi.org/10.1016/j.cpr.2019.101810

Goodman, F. R., Peckham, A. D., Kneeland, E. T., Choate, A. M., Daniel, K. E., Beard, C., & Björgvinsson, T. (2023). How does emotion regulation change during psychotherapy? A daily diary study of adults in a transdiagnostic partial hospitalization program. *Journal of Consulting and Clinical Psychology*, *91*(12), 731. https://doi.org/10.1037/ccp0000838.

Gois, C., Dias, V. V., Carmo, I., Duarte, R., Ferro, A., Santos, A. L., Sousa, F., & Barbosa, A. (2014). Treatment response in type 2 diabetes patients with major depression. *Clinical Psychology & Psychotherapy*, 21(1), 39–48. https://doi.org/10.1002/cpp.1817

Goldwin, M. & Behar, E. (2012): Concreteness of Idiographic Periods of Worry and Depressive Rumination. *Cogn Ther Res* 36 (6), 840–846. https://doi.org/10.1007/s10608-011-9428-1.

Gorin, A. A., & Stone, A. A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. *Handbook of Health Psychology, 23*, 405–413.

Gotlib, I. H., & Joormann, J. (2010). Cognition and Depression: Current Status and Future Directions. *Annu. Rev. Clin. Psychol.*, *6*(1), 285–312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Gratch, I., Choo, T. H., Galfalvy, H., Keilp, J. G., Itzhaky, L., Mann, J. J., ... & Stanley, B. (2021). Detecting suicidal thoughts: The power of ecological momentary assessment. *Depression and anxiety*, *38*(1), 8-16. https://doi.org/10.1002/da.23043.

Grawe, K. (1995). Grundriss einer allgemeinen Psychotherapie. *Psychotherapeut, 40*(3), 130–145.

Greenberg, L. S. (2004). Emotion–focused therapy. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 11(1), 3–16. https://doi.org/10.1002/cpp.388

Greenberg, M. S., & Beck, A. T. (1989). Depression versus anxiety: a test of the content-specificity hypothesis. *Journal of Abnormal Psychology, 98*(1), 9. https://doi.org/10.1037/0021-843X.98.1.9

Gutiérrez-Rojas, L., Porras-Segovia, A., Dunne, H., Andrade-González, N., & Cervilla, J. A. (2020). Prevalence and correlates of major depressive disorder: a systematic review. *Brazilian Journal of Psychiatry*, 42, 657-672. https://dx.doi.org/10.1590/1516-4446-2020-0650.

Hall, M., Scherner, P. V., Kreidel, Y., & Rubel, J. A. (2021). A systematic review of momentary assessment designs for mood and anxiety symptoms. *Frontiers in Psychology*, *12*, 642044. https://doi.org/10.3389/fpsyg.2021.642044.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological methods*, *20*(1), 102. http://dx.doi.org/10.1037/a0038889.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747.

Harnas, S. J., Knoop, H., Booij, S. H., & Braamse, A. M. (2021). Personalizing cognitive behavioral therapy for cancer-related fatigue using ecological momentary assessments followed by automated individual time series analyses: a case report series. *Internet Interventions*, *25*, 100430. https://doi.org/10.1016/j.invent.2021.100430.

References

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R'* (Version 0.0.9000) [Computer software]. http://dmetar.protectlab.org/.

Hartmann, R., Schmidt, F. M., Sander, C., & Hegerl, U. (2019). Heart rate variability as indicator of clinical state in depression. *Frontiers in psychiatry*, *9*, 735. https://doi.org/10.3389/fpsyt.2018.00735.

Hautzinger, M., Keller, F., & Kühner, C. (2009). BDI-II. Beck-Depressions-Inventar. Revision. 2, Auflage. Frankfurt: Pearson Assessment.

Heo, J. Y., Kim, K., Fava, M., Mischoulon, D., Papakostas, G. I., Kim, M. J., ... & Jeon, H. J. (2017). Effects of smartphone use with and without blue light at night in healthy adults: A randomized, double-blind, cross-over, placebo-controlled comparison. *Journal of psychiatric research*, *87*, 61-70. https://doi.org/10.1016/j.jpsychires.2016.12.010.

Herpertz, S. C., & Schramm, E. (Eds.). (2022). *Modulare psychotherapie: ein mechanismus-basiertes, personalisiertes vorgehen*. Klett-Cotta.

Hjartarson, K. H., Snorrason, I., Bringmann, L. F., Ögmundsson, B. E., & Ólafsson, R. P. (2021). Do daily mood fluctuations activate ruminative thoughts as a mental habit? Results from an ecological momentary assessment study. *Behaviour Research and Therapy*, *140*, 103832. https://doi.org/10.1016/j.brat.2021.103832.

Hofmann, S. G., Curtiss, J., & McNally, R. J. (2016). A complex network perspective on clinical science. *Perspectives on Psychological Science*, *11*(5), 597-605. https://doi.org/10.1177/17456916166392.

Hong, R. Y. (2007). Worry and rumination: Differential associations with anxious and depressive symptoms and coping behavior. *Behaviour research and therapy*, *45*(2), 277-290. https://doi.org/10.1016/j.brat.2006.03.006.

Hornstein, S., Zantvoort, K., Lueken, U., Funk, B., & Hilbert, K. (2023). Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. *Frontiers in digital health*, *5*, 1170002. https://doi.org/10.3389/fdgth.2023.1170002.

Huibers, M. J. H., Cohen, Z. D., Lemmens, Lotte H. J. M., Arntz, A., Peeters, Frenk P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLOS ONE*, 10(11), e0140771. https://doi.org/10.1371/journal.pone.0140771.

Huibers, M. J., Lorenzo-Luaces, L., Cuijpers, P., & Kazantzis, N. (2021). On the road to personalized psychotherapy: A research agenda based on cognitive behavior therapy for depression. *Frontiers in Psychiatry*, *11*, 607508. https://doi.org/10.3389/fpsyt.2020.607508.

Hung, G. C. L., Yang, P. C., Wang, C. Y., & Chiang, J. H. (2015, December). A smartphone-based personalized activity recommender system for patients with depression. In *Proceedings of the 5th EAI international conference on wireless mobile communication and healthcare* (pp. 253-257). https://doi.org/10.4108/eai.14-10-2015.2261655.

Hunnicutt-Ferguson, K., Hoxha, D., & Gollan, J. (2012). Exploring sudden gains in behavioral activation therapy for Major Depressive Disorder. *Behaviour Research and Therapy*, 50(3), 223–230. https://doi.org/10.1016/j.brat.2012.01.005

Ingram, R. E., & Hollon, S. D. (1986). Cognitive therapy for depression from an information processing perspective. In R. E. Ingram (Ed.), *Information processing approaches to clinical psychology* (pp. 255–281). San Diego, CA: Academic Press. https://psycnet.apa.org/record/1986-98710-013.

Jacobson, N. C. (2020). Compliance thresholds in intensive longitudinal data: Worse than listwise deletion: Call for action. *Society for Ambulatory Assessment*, Melbourne, Australia. Retrieved January 26, 2024, from: https://www.nicholasjacobson.com/talk/saa2020_compliance_thresholds/

Joormann, J., Dkane, M., & Gotlib, I. H. (2006). Adaptive and maladaptive components of rumination? Diagnostic specificity and relation to depressive biases. *Behavior therapy*, *37*(3), 269-280. https://doi.org/10.1016/j.beth.2006.01.002.

Kardum, I., & Daskijević, K. T. (2001). Absolute and relative accuracy in the retrospective estimate of positive and negative mood. *European Journal of Psychological Assessment*, *17*(1), 69. https://doi.org/10.1027/1015-5759.17.1.69

References

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), e1858. https://10.2196/jmir.1858.

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, 73(1), 1–2. https://doi.org/10.1037/amp0000263

Keller, M. B. (2003). Past, present, and future directions for defining optimal treatment outcome in depression: Remission and beyond. *JAMA: Journal of the American Medical Association*, 289(23), 3152–3160. https://doi.org/10.1001/jama.289.23.3152

Kessler, R. C., Sampson, N. A., Berglund, P., Gruber, M. J., Al-Hamzawi, A., Andrade, L., ... & Wilcox, M. A. (2015). Anxious and non-anxious major depressive disorder in the World Health Organization World Mental Health Surveys. *Epidemiology and psychiatric sciences*, 24(3), 210-226. https://doi.org/10.1017/S2045796015000189.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., ... & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and psychiatric sciences*, *26*(1), 22-36. https://doi.org/10.1017/S2045796016000020.

Kheiri, K., & Karimi, H. (2023). Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*. https://doi.org/10.48550/arXiv.2307.10234.

Kidwell, K. M., & Almirall, D. (2023). Sequential, multiple assignment, randomized trial designs. *Jama*, *329*(4), 336-337. https://doi.org/10.1001/jama.2022.24324.

Kim, T., Bae, S., Kim, H. A., Lee, S. W., Hong, H., Yang, C., & Kim, Y. H. (2024, May). MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-20). https://doi.org/10.1145/3613904.3642937.

Kircanski, K., Thompson, R. J., Sorenson, J., Sherdell, L., & Gotlib, I. H. (2015). Rumination and Worry in Daily Life: Examining the Naturalistic Validity of Theoretical Constructs. *Clinical Psychological Science: A Journal of the Association for Psychological Science*, 3(6), 926–939. https://doi.org/10.1177/2167702614566603.

Kircanski, K., Thompson, R. J., Sorenson, J., Sherdell, L., & Gotlib, I. H. (2018). The everyday dynamics of rumination and worry: Precipitant events and affective consequences. *Cognition and Emotion*, 32(7), 1424-1436. https://doi.org/10.1080/02699931.2017.1278679.

Kirchberger, I., Braitmayer, K., Coenen, M., Oberhauser, C., & Meisinger, C. (2014). Feasibility and psychometric properties of the German 12-item WHO Disability Assessment Schedule (WHODAS 2.0) in a population-based sample of patients with myocardial infarction from the MONICA/KORA myocardial infarction registry. *Population Health Metrics*, *12*(1), 1–13. https://doi.org/10.1186/s12963-014-0027-8.

Koch, C., Wilhelm, M., Salzmann, S., Rief, W., & Euteneuer, F. (2019). A meta-analysis of heart rate variability in major depression. *Psychological medicine*, *49*(12), 1948-1957. https://doi.org/10.1017/S0033291719001351.

König, H., König, H. H., & Konnopka, A. (2020). The excess costs of depression: a systematic review and meta-analysis. *Epidemiology and psychiatric sciences*, 29, e30. https://doi.org/10.1017/S2045796019000180.

Köper, M., & Im Schulte Walde, S. (2016). Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. *In LREC*.

Kopf-Beck, J., Müller, C. L., Tamm, J., Fietz, J., Rek, N., Just, L., Spock, Z. I., Weweck, K., Takano, K., Rein, M., Keck, M. E., & Egli, S. (2024). Effectiveness of Schema Therapy versus Cognitive Behavioral Therapy versus Supportive Therapy for Depression in Inpatient and Day Clinic Settings: A Randomized Clinical Trial. *Psychother Psychosom*, 93(1), 24–35. https://doi.org/10.1159/000535492.

Kopf-Beck, J., Zimmermann, P., Egli, S., Rein, M., Kappelmann, N., Fietz, J., Tamm, J., Rek, K., Lucae, S., & Brem, A.-K. (2020). Schema therapy versus cognitive behavioral therapy versus individual supportive therapy for depression in an inpatient and day clinic setting: study protocol of the OPTIMA-RCT. *BMC Psychiatry*, 20, 1–19. https://doi.org/10.1186/s12888-020-02880-x.

Krippendorff, K. (2019). Content analysis: An introduction to its methodology (Fourth Edition). Los Angeles: *SAGE*. https://dx.doi.org/10.4135/9781071878781.

References

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine, 16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*(1), 537-567. https://doi.org/10.1146/annurev.psych.50.1.537.

Kung, S., Alarcon, R. D., Williams, M. D., Poppe, K. A., Moore, M. J., & Frye, M. A. (2013). Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an integrated mood disorders practice. *Journal of affective disorders*, *145*(3), 341-343. https://doi.org/10.1016/j.jad.2012.08.017.

Lamers, F., van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., van Balkom, A. J., ... & Penninx, B. W. (2011). Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands Study of Depression and Anxiety (NESDA). *The Journal of clinical psychiatry*, 72(3), 3397. https://doi.org/10.4088/JCP.10m06176blu.

Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024). Depression Detection on Social Media with Large Language Models. *arXiv preprint arXiv:2403.10750.* https://doi.org/10.48550/arXiv.2403.10750.

Levine, L. J., Schmidt, S., Kang, H. S., & Tinti, C. (2012). Remembering the silver lining: Reappraisal and positive bias in memory for emotion. *Cognition & emotion*, *26*(5), 871-884. https://doi.org/10.1080/02699931.2011.625403.

Li, F., Jörg, F., Merkx, M. J.M., & Feenstra, T. (2023). Early symptom change contributes to the outcome prediction of cognitive behavioral therapy for depression patients: A machine learning approach. *Journal of Affective Disorders*, 334, 352–357. https://doi.org/10.1016/j.jad.2023.04.111.

Long, J. A. (2022). jtools: Analysis and Presentation of Social Scientific Data [Computer software]. Retrieved from https://cran.r-project.org/package=jtools.

Lorenzo-Luaces, L., Peipert, A., De Jesus Romero, R., Rutter, L. A., & Rodriguez-Quintana, N. (2021). Personalized medicine and cognitive behavioral therapies for depression: Small effects, big problems, and bigger data. *International Journal of Cognitive Therapy*, 14, 59-85. https://doi.org/10.1007/s41811-020-00094-3.

Lutz, W., Arndt, A., Rubel, J., Berger, T., Schröder, J., Späth, C., Meyer, B., Greiner, W., Gräfe, V., Hautzinger, M., Fuhr, K., Rose, M., Nolte, S., Löwe, B., Hohagen, F., Klein, J. P., & Moritz, S. (2017). Defining and predicting patterns of early response in a web-based intervention for depression. *Journal of Medical Internet Research*, 19(6), 40–55. https://doi.org/10.2196/jmir.7367

Lyubomirsky, Sonja; Nolen-Hoeksema, Susan (1995): Effects of self-focused rumination on negative thinking and interpersonal problem solving. *Journal of personality and social psychology* 69 (1), 176 - 190. https://doi.org/10.1037/0022-3514.69.1.176.

MacIsaac, A., Mushquash, A. R., & Wekerle, C. (2023). Writing yourself well: dispositional self-reflection moderates the effect of a smartphone app-based journaling intervention on psychological wellbeing across time. *Behaviour Change*, *40*(4), 297-313. https://doi.org/10.1017/bec.2022.24.

Malhi, G. S., Hamilton, A., Morris, G., Mannie, Z., Das, P., & Outhred, T. (2017). The promise of digital mood tracking technologies: are we heading on the right track? *Evidence Based Mental Health, 20*(4), 102. https://doi.org/10.1136/eb-2017-102757

McGowan, S. K., Stevens, E. S., Behar, E., Judah, M. R., Mills, A. C., & Grant, D. M. (2017). Concreteness of idiographic worry and anticipatory processing. *Journal of Behavior Therapy and Experimental Psychiatry*, *54*, 195-203. https://doi.org/10.1016/j.jbtep.2016.08.005.

McGrath, J. J., Lim, C. C. W., Plana-Ripoll, O., Holtz, Y., Agerbo, E., Momen, N. C., ... & De Jonge, P. (2020). Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiology and Psychiatric Sciences*, 29, e153. https://doi.org/10.1017/S2045796020000633.

McKay, D. E., Abramowitz, J. S., & Taylor, S. E. (2010). Cognitive-behavioral therapy for refractory cases: Turning failure into success. *American Psychological Association*. https://doi.org/10.1037/12070-000.

Mekonen, T., Chan, G. C., Connor, J. P., Hides, L., & Leung, J. (2021). Estimating the global treatment rates for depression: a systematic review and meta-analysis. *Journal of Affective Disorders*, 295, 1234-1242. https://doi.org/10.1016/j.jad.2021.09.038.

References

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., ... & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist, 56*(2), 128. http://dx.doi.org/10.1037/0003-066X.56.2.128.

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus and negative affect: an experience sampling study. *Journal of abnormal psychology*, 117(2), 314. https://doi.org/10.1037/0021-843X.117.2.314.

Monteregge, S., Tsagkalidou, A., Cuijpers, P., & Spinhoven, P. (2020). The effects of different types of treatment for anxiety on repetitive negative thinking: A meta-analysis. *Clinical Psychology: Science and Practice*, 27(2), 110. https://doi.org/10.1111/cpsp.12316.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry: The Journal of Mental Science*, 134(4), 382–389. https://doi.org/10.1192/bjp.134.4.382.

Moore, R. C., Depp, C. A., Wetherell, J. L., & Lenze, E. J. (2016). Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. *Journal of Psychiatric Research*, 75, 116–123. https://doi.org/10.1016/j.jpsychires.2016.01.011

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. *Psychological Bulletin, 128*(4), 638–662. https://doi.org/10.1037/0033-2909.128.4.638.

Moshe, I., Terhorst, Y., Philippi, P., Domhardt, M., Cuijpers, P., Cristea, I., ... & Sander, L. B. (2021). Digital interventions for the treatment of depression: A meta-analytic review. *Psychological bulletin*, *147*(8), 749. https://doi.org/10.1037/bul0000334.

Moskowitz, D. S., & Young, S. N. (2006). Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience, 31*(1), 13. https://pubmed.ncbi.nlm.nih.gov/16496031/.

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. https://doi.org/10.1002/wps.20513.

Navarro, D. J. (2015). *Learning Statistics with R: A tutorial for psychology students and other beginners*. (Version 0.5.2) [Computer software]. University of New South Wales. https://CRAN.R-project.org/package=lsr.

Nelson, J., Klumparendt, A., Doebler, P., & Ehring, T. (2020). Everyday emotional dynamics in major depression. Emotion, 20(2), 179.

Nezu, A. M., Nezu, C. M., & D'Zurilla, T. J. (2013). Problem-solving therapy: A treatment manual: A Treatment Manual. New York: *Springer Publishing Company*.

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology*, 100(4), 569–582. https://doi.org/10.1037//0021-843x.100.4.569.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology*, 109(3), 504–511. https://doi.org/10.1037//0021-843x.109.3.504.

Nolen-Hoeksema, S., Morrow, J., & Fredrickson, B. L. (1993). Response styles and the duration of episodes of depressed mood. *Journal of Abnormal Psychology*, 102(1), 20–28. https://doi.org/10.1037//0021-843x.102.1.20.

Nolen-Hoeksema, Susan; Morrow, Jannay (1993): Effects of rumination and distraction on naturally occurring depressed mood. *Cogn Emot* 7 (6), 561 - 570. https://doi.org/10.1080/02699939308409206.

Nolen-Hoeksema, S., & Watkins, E. R. (2011). A Heuristic for Developing Transdiagnostic Models of Psychopathology: Explaining Multifinality and Divergent Trajectories. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 6(6), 589–609. https://doi.org/10.1177/1745691611419672.

Nolen-Hoeksema, Susan; Wisco, Blair E.; Lyubomirsky, Sonja (2008): Rethinking rumination. *Perspect Psychol Sci* 3 (5), 400 - 424. https://doi.org/10.1111/j.1745-6924.2008.00088.x.

References

Olatunji, B. O., Knowles, K. A., Cox, R. C., & Cole, D. A. (2023). Linking repetitive negative thinking and insomnia symptoms: A longitudinal trait-state model. *Journal of Anxiety Disorders*, *97*, 102732. https://doi.org/10.1016/j.janxdis.2023.102732.

Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR mHealth and uHealth, 9(7), e26540. https://doi.org/10.2196/26540.

OpenAI. (2024). *ChatGPT (GPT-4)* [Large language model]. https://chat.openai.com/

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion (Washington, D.C.)*, *6*(3), 383. https://doi.org/10.1037/1528-3542.6.3.383

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic Inquiry and Word Count: LIWC2015 [Computer software]. Austin, TX: Pennebaker Conglomerates: Pennebaker Conglomerates. Retrieved from www.LIWC.net.

Peipert, A., Krendl, A. C., & Lorenzo-Luaces, L. (2022). Waiting lists for psychotherapy and provider attitudes toward low-intensity treatments as potential interventions: Survey study. *JMIR formative research*, *6*(9), e39787. https://doi.org/10.2196/39787.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). *nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1.160* [Computer software]. https://CRAN.R-project.org/package=nlme.

Pizzoli, S. F., Marzorati, C., Gatti, D., Monzani, D., Mazzocco, K., & Pravettoni, G. (2021). A meta-analysis on heart rate variability biofeedback and depressive symptoms. *Scientific reports*, *11*(1), 6650. https://doi.org/10.1038/s41598-021-86149-7.

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological assessment*, *22*(2), 203. https://doi.org/10.1037/a0018679.

R Core Team. (2020). *R: A language and environment for statistical*. R Foundation for Statistical Computing. https://www.R-project.org/.

Revelle, W. (2022). *psych: Procedures for psychological, psychometric, and personality research. R package version 2.2. 3* [Computer software]. https://CRAN.R-project.org/package=psych.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395. https://doi.org/10.1037/pas0000754.

Richards, D. A. (2012). Stepped Care: A Method to Deliver Increased Access to Psychological Therapies. Canadian Journal of Psychiatry. *Revue Canadienne De Psychiatrie*, 57(4), 210–215. https://doi.org/10.1177/070674371205700403.

Rimpler, A., Siepe, B. S., Rieble, C. L., Proppert, R. K., & Fried, E. I. (2024). Introducing FRED: Software for generating feedback reports for ecological momentary assessment data. *Administration and Policy in Mental Health and Mental Health Services Research*, 1-11. https://doi.org/10.1007/s10488-023-01324-4.

Roberts, H., Jacobs, R. H., Bessette, K. L., Crowell, S. E., Westlund-Schreiner, M., Thomas, L., . . . Watkins, E. R. (2021). Mechanisms of rumination change in adolescent depression (RuMeChange): Study protocol for a randomised controlled trial of rumination-focused cognitive behavioural therapy to reduce ruminative habit and risk of depressive relapse in high-ruminating adolescents. *BMC Psychiatry*, 21(1), 206. https://doi.org/10.1186/s12888-021-03193-3.

Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry*. 1988 Dec;45(12):1069–77. https://doi.org/10.1001/archpsyc.1988.01800360017003.

Rosenkranz, T., Takano, K., Watkins, E. R., & Ehring, T. (2020). Assessing repetitive negative thinking in daily life: Development of an ecological momentary assessment paradigm. *PLOS ONE, 15*(4), e0231783. https://doi.org/10.1371/journal.pone.0231783.

Rubel, J., Lutz, W., Kopta, S. M., Köck, K., Minami, T., Zimmermann, D., & Saunders, S. M. (2015). Defining early positive response to psychotherapy: An empirical comparison between clinically significant change criteria and growth mixture modeling. *Psychological Assessment*, 27(2), 478–488. https://doi.org/10.1037/pas0000060.

References

Ruscio, A. M., Gentes, E. L., Jones, J. D., Hallion, L. S., Coleman, E. S., & Swendsen, J. (2015). Rumination predicts heightened responding to stressful life events in major depressive disorder and generalized anxiety disorder. *Journal of abnormal psychology*, 124(1), 17. https://doi.org/10.1037/abn0000025.

Rush, A. J., Kraemer, H. C., Sackeim, H. A., Fava, M., Trivedi, M. H., Frank, E., ... & Schatzberg, A. F. (2006). Report by the ACNP Task Force on response and remission in major depressive disorder. *Neuropsychopharmacology*, 31(9), 1841-1853. https://doi.org/10.1038/sj.npp.1301131

Saberi, S., Ahmadi, R., Khakpoor, S., Pirzeh, R., Hasani, M., Moradveisi, L., & Saed, O. (2024). Comparing the effectiveness of behavioral activation in group vs. self-help format for reducing depression, repetitive thoughts, and enhancing performance of patients with major depressive disorder: a randomized clinical trial. *BMC psychiatry*, 24(1), 516. https://doi.org/10.1186/s12888-024-05973-z.

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., ... & Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, 398(10312), 1700-1712. https://doi.org/10.1016/S0140-6736(21)02143-7.

Schaffer, A., Kreindler, D., Reis, C., & Levitt, A. J. (2013). Use of mental health telemetry to enhance identification and predictive value of early changes during augmentation treatment of major depression. *Journal of Clinical Psychopharmacology*, 33(6), 775–781. https://doi.org/10.1097/JCP.0b013e31829e8359.

Schibbye, P., Ghaderi, A., Ljótsson, B., Hedman, E., Lindefors, N., Rück, C., & Kaldo, V. (2014). Using Early Change to Predict Outcome in Cognitive Behaviour Therapy: Exploring Timeframe, Calculation Method, and Differences of Disorder-Specific versus General Measures. *PLOS ONE*, 9(6), e100614. https://doi.org/10.1371/journal.pone.0100614.

Schrijvers, D., Hulstijn, W., & Sabbe, B. G. (2008). Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool. *Journal of affective disorders*, *109*(1-2), 1-20. https://doi.org/10.1016/j.jad.2007.10.019.

Schuster, R., Schreyer, M. L., Kaiser, T., Berger, T., Klein, J. P., Moritz, S., ... & Trutschnig, W. (2020). Effects of intense assessment on statistical power in randomized controlled

trials: Simulation study on depression. *Internet Interventions*, 20, 100313. https://doi.org/10.1016/j.invent.2020.100313.

Schramm, E., Elsaesser, M., Jenkner, C., Hautzinger, M., & Herpertz, S. C. (2024). Algorithm-based modular psychotherapy vs. cognitive-behavioral therapy for patients with depression, psychiatric comorbidities and early trauma: a proof-of-concept randomized controlled trial. *World Psychiatry*, *23*(2), 257-266. https://doi.org/10.1002/wps.21204.

Segerstrom, S. C., Stanton, A. L., Alden, L. E., & Shortridge, B. E. (2003). A multidimensional structure for repetitive thought: What's on your mind, and how, and how much? *Journal of Personality and Social Psychology*, 85(5), 909–921. https://doi.org/10.1037/0022-3514.85.5.909.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415.

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior research methods*, *46*, 41-54. https://doi.org/10.3758/s13428-013-0353-y.

Simmonds-Buckley, M., Catarino, A., & Delgadillo, J. (2021). Depression subtypes and their response to cognitive behavioral therapy: A latent transition analysis. *Depression and anxiety*, *38*(9), 907-916. https://doi.org/10.1002/da.23161.

Shin, D., Kim, H., Lee, S., Cho, Y., & Jung, W. (2024). Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study. *Journal of Medical Internet Research*, *26*, e54617. https://doi.org/10.2196/54617.

Snijders, T. A. B., & Bosker, R. (2011). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). Los Angeles: Sage Publications Ltd.

Snippe, E., Doornbos, B., Schoevers, R. A., Wardenaar, K. J., & Wichers, M. (2021). Individual and common patterns in the order of symptom improvement during

outpatient treatment for major depression. *Journal of affective disorders*, *290*, 81-88. https://doi.org/10.1016/j.jad.2021.04.097.

Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, *30*(5), 1662-1671. https://doi.org/10.1177/10731911221113563.

Spinhoven, Philip; Klein, Nicola; Kennis, Mitzy; Cramer, Angélique O. J.; Siegle, Greg; Cuijpers, Pim et al. (2018): The effects of cognitive-behavior therapy for depression on repetitive negative thinking. A meta-analysis. *Behaviour Research and Therapy*, 106, 71–85. https://doi.org/10.1016/j.brat.2018.04.002.

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., . . . Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *Npj Mental Health Research*, 3(1), 12. https://doi.org/10.1038/s44184-024-00056-z.

Starr, L. R., Hershenberg, R., Li, Y. I., & Shaw, Z. A. (2017). When Feelings Lack Precision: Low Positive and Negative Emotion Differentiation and Depressive Symptoms in Daily Life. *Clinical Psychological Science: A Journal of the Association for Psychological Science*, 5(4), 613–631. https://doi.org/10.1177/2167702617694657.

Statista (2023). Ranking der Länder mit der höchsten durchschnittlichen Dauer der mobilen Internetnutzung weltweit im Jahr 2023. Retrieved August 24, 2024. from: https://de.statista.com/statistik/daten/studie/809553/umfrage/taegliche-dauer-der-mobilen-internetnutzung-nach-laendern-weltweit.

Statista (2024). Number of smartphone users worldwide from 2014 to 2029 (in millions). Retrieved August 24, 2024. from: https://www.statista.com/forecasts/1143723/smartphone-users-in-the-world.

Stöber, J. (1998). Worry, problem elaboration and suppression of imagery: the role of concreteness. *Behaviour Research and Therapy*, 36(7-8), 751–756. https://doi.org/10.1016/S0005-7967(98)00027-8.

Stöber, J. (2000). Worry, Thoughts, and Images: A New Conceptualization. In U. von Hecker, S. Dutke, & G. Sedek (Eds.), Generative mental processes and cognitive resources:

Integrative research on adaptation and control (pp. 223–244). *Dordrecht: Springer-Science+Business Media*, B.V. https://doi.org/10.1007/978-94-011-4373-8_9.

Stöber, J., & Borkovec, T. D. (2002). Reduced Concreteness of Worry in Generalized Anxiety Disorder: Findings from a Therapy Study. *Cognitive Therapy and Research*, 26(1), 89–96. https://doi.org/10.1023/A:1013845821848.

Stöber, J., Tepperwien, S., & Staak, M. (2000). Worrying leads to reduced concreteness of problem elaborations: Evidence for the avoidance theory of worry. *Anxiety, Stress & Coping* 13 (3), 217–227. https://doi.org/10.1080/10615800008549263.

Stuart, A. L., Pasco, J. A., Jacka, F. N., Brennan, S. L., Berk, M., & Williams, L. J. (2014). Comparison of self-report and structured clinical interview in the identification of depression. *Comprehensive psychiatry*, *55*(4), 866-869. https://doi.org/10.1016/j.comppsych.2013.12.019.

Szegedi, A., Jansen, W. T., van Willigenburg, A. P. P., van der Meulen, E., Stassen, H. H., & Thase, M. E. (2009). Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. *J Clin Psychiatry*, 70(3), 344-53. https://doi.org/10.4088/jcp.07m03780.

Takano, K., & Tanno, Y. (2010). Concreteness of thinking and self-focus. *Consciousness and Cognition*, 19(1), 419–425. https://doi.org/10.1016/j.concog.2009.11.010.

Takano, K., & Tanno, Y. (2011). Diurnal variation in rumination. *Emotion* (Washington, D.C.), 11(5), 1046–1058. https://doi.org/10.1037/a0022757.

Tamm, J., Takano, K., Just, L., Ehring, T., Rosenkranz, T. & Kopf-Beck, J. (2024). Ecological Momentary Assessment versus Weekly Questionnaire Assessment of Change in Depression. *Depression and Anxiety*, 2024(1), 9191823. https://doi.org/10.1155/2024/9191823.

Targum, S. D. (2020). Baseline reliability and early response in clinical trials of major depressive disorder. *Medical Research Archives*, 8(10). https://doi.org/10.18103/mra.v8i10.2241.

Targum, S. D., Sauder, C., Evans, M., Saber, J. N., & Harvey, P. D. (2021). Ecological momentary assessment as a measurement tool in depression trials. *Journal of Psychiatric Research, 136*, 256–264. https://doi.org/10.1016/j.jpsychires.2021.02.012.

Taylor, M. M., & Snyder, H. R. (2021). Repetitive negative thinking shared across rumination and worry predicts symptoms of depression and anxiety. *Journal of Psychopathology and Behavioral Assessment*, 43(4), 904-915. https://doi.org/10.1007/s10862-021-09898-9.

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Warden, D., McKinney, W., Downing, M., Berman, S. R., Farabaugh, A., Luther, J. F., & Nierenberg, A. A. (2006). Factors associated with health-related quality of life among outpatients with major depressive disorder: a STAR* D report. *The Journal of Clinical Psychiatry*, 67(2), 185–195. https://doi.org/10.4088/jcp.v67n0203.

Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *1939-134X*. https://doi.org/10.1037/a0017653.

Trull, T. J., and Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: a review of recommended reporting guidelines and current practices. *J. Abnorm. Psychol.* 129, 56–63. https://doi.org/10.1037/abn0000473.

van Beljouw, I., Verhaak, P., Prins, M., Cuijpers, P., Penninx, B., & Bensing, J. (2010). Reasons and determinants for not receiving treatment for common mental disorders. *Psychiatric Services*, *61*(3), 250-257. https://doi.org/10.1176/ps.2010.61.3.250.

van Dijk, D. A., Meijer, R. M., van den Boogaard, T. M., Spijker, J., Ruhé, H. G., & Peeters, F. P. M. L. (2023). Worse off by waiting for treatment? The impact of waiting time on clinical course and treatment outcome for depression in routine care. *Journal of affective disorders*, *322*, 205-211. https://doi.org/10.1016/j.jad.2022.11.011.

van Genugten, C. R., Schuurmans, J., Lamers, F., Riese, H., Penninx, B. W., Schoevers, R. A., ... & Smit, J. H. (2020). Experienced burden of and adherence to smartphone-based ecological momentary assessment in persons with affective disorders. *Journal of clinical medicine*, 9(2), 322. https://doi.org/10.3390/jcm9020322.

van Genugten, C. R., Schuurmans, J., Hoogendoorn, A. W., Araya, R., Andersson, G., Baños, R. M., ... & Riper, H. (2022). A Data-Driven Clustering Method for Discovering Profiles in the Dynamics of Major Depressive Disorder Using a Smartphone-Based

Ecological Momentary Assessment of Mood. *Frontiers in psychiatry*, *13*, 755809. https://doi.org/10.3389/fpsyt.2022.755809.

van Straten, A., Hill, J., Richards, D. A., & Cuijpers, P. (2015). Stepped care treatment delivery for depression: a systematic review and meta-analysis. *Psychological medicine*, *45*(2), 231-246. https://doi.org/10.1017/S0033291714000701.

von Klipstein, L., Riese, H., van der Veen, D. C., Servaas, M. N., & Schoevers, R. A. (2020). Using person-specific networks in psychotherapy: challenges, limitations, and how we could use them anyway. *BMC medicine*, *18*, 1-8. https://doi.org/10.1186/s12916-020-01818-0.

Wahl, K., Ehring, T., Kley, H., Lieb, R., Meyer, A., Kordon, A., ... & Schönfeld, S. (2019). Is repetitive negative thinking a transdiagnostic process? A comparison of key processes of RNT in depression, generalized anxiety disorder, obsessive-compulsive disorder, and community controls. *Journal of behavior therapy and experimental psychiatry*, 64, 45-53. https://doi.org/10.1016/j.jbtep.2019.02.006.

Wallsten, D., Norell, A., Anniko, M., Eriksson, O., Lamourín, V., Halldin, I., . . . Tillfors, M. (2023). Treatment of worry and comorbid symptoms within depression, anxiety, and insomnia with a group-based rumination-focused cognitive-behaviour therapy in a primary health care setting: A randomised controlled trial. *Frontiers in Psychology*, 14, 1196945. https://doi.org/10.3389/fpsyg.2023.1196945.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin*, 134(2), 163–206. https://doi.org/10.1037/0033-2909.134.2.163.

Watkins, E. R., & Moberly, N. J. (2009). Concreteness training reduces dysphoria: A pilot proof-of-principle study. *Behaviour Research and Therapy*, 47(1), 48–53. https://doi.org/10.1016/j.brat.2008.10.014.

Watkins, E. R., & Moulds, M. L. (2007). Reduced concreteness of rumination in depression: A pilot study. *Personality and Individual Differences*, 43(6), 1386–1395. https://doi.org/10.1016/j.paid.2007.04.007.

Watkins, E. D., & Moulds, M. (2005a). Distinct modes of ruminative self-focus: impact of abstract versus concrete rumination on problem solving in depression. *Emotion*, *5*(3), 319. https://doi.org/10.1037/1528-3542.5.3.319.

Watkins, E. R., Moulds, M. L., & Mackintosh, B. (2005b). Comparisons between rumination and worry in a non-clinical population. *Behaviour Research and Therapy*, 43(12), 1577–1585. https://doi.org/10.1016/j.brat.2004.11.008.

Watkins, E.R., & Roberts, H. (2020). Reflecting on rumination: Consequences, causes, mechanisms and treatment of rumination. *Behaviour Research and Therapy, 127*, 103573 . https://doi.org/10.1016/j.brat.2020.103573.

Watkins, E. R., Taylor, R. S., Byng, R., Baeyens, C., Read, R. J., Pearson, K. A., & Watson, L. (2012). Guided self-help concreteness training as an intervention for major depression in primary care: A Phase II randomized controlled trial. *Psychological Medicine*, 42(7), 1359–1371. https://doi.org/10.1017/S0033291711002480.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063.

Werner-Seidler, A., & Moulds, M. L. (2012). Mood repair and processing mode in depression. *Emotion* (Washington, D.C.), 12(3), 470–478. https://doi.org/10.1037/a0025984.

Wichers, M., Simons, C. J. P., Kramer, I. M. A., Hartmann, J. A., Lothmann, C., Myin-Germeys, I., ... & Van Os, J. (2011). Momentary assessment technology as a tool to help patients with depression help themselves. *Acta psychiatrica scandinavica*, *124*(4), 262-272. https://doi.org/10.1111/j.1600-0447.2011.01749.x.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. Retrieved from https://ggplot2.tidyverse.org.

Wirz-Justice, A. (2008). Diurnal variation of depressive symptoms. *Dialogues in Clinical Neuroscience, 10*(3), 337–343. https://doi.org/10.31887/DCNS.2008.10.3/awjustice.

Wirtz, D., Kruger, J., Scollon, C. N., & Diener, E. (2003). What to do on spring break? The role of predicted, on-line, and remembered experience in future choice. *Psychological Science, 14*(5), 520–524. https://doi.org/10.1111/1467-9280.03455.

Wittchen HU, Lachner G, Wunderlich U, Pfister H. Test-retest reliability of the computerized DSM-IV version of the Munich-Composite International Diagnostic Interview (M-

CIDI). *Soc Psychiatry Psychiatr Epidemiol*. 1998 Oct 1;33(11):568–78. https://doi.org/10.1007/s001270050095.

World Health Organization. (2017, March 30). *Depression: Let's talk, says WHO, as depression tops list of causes of ill health*. Retrieved September 2, 2024, from https://www.who.int/news/item/30-03-2017--depression-let-s-talk-says-who-as-depression-tops-list-of-causes-of-ill-health.

World Health Organization. (2022). ICD-11: International classification of diseases (11th revision). https://icd.who.int/.

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*. https://doi.org/10.48550/arXiv.2304.03347.

Yıldırım, S. (2023). The Challenge of Self-diagnosis on Mental Health Through Social Media: A Qualitative Study. In *Computational Methods in Psychiatry* (pp. 197-213). *Singapore: Springer Nature Singapore*. https://doi.org/10.1007/978-981-99-6637-0_10.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. http://dx.doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3.

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2006). *Schema therapy: A practitioner's guide*. Guilford press.

Zetsche, U., Bürkner, P. C., & Renneberg, B. (2019). Future expectations in clinical depression: Biased or realistic?. *Journal of Abnormal Psychology*, *128*(7), 678. https://doi.org/10.1037/abn0000452 .

List of Figures

Study I

Figure 2.1. Data Exclusion Flow Diagram 42
Figure 2.2. Comparison of EMA versus WQA Data in the Time Series 50

Study II

Figure 3.1. Flow Diagram 68
Figure 3.2. Receiver Operating Characteristic (ROC) Plots for the Prediction of Treatment Response after 7 weeks through Early Improvement in Depressive Symptoms measured by Ecological Momentary Assessment or Weekly Questionnaire Assessment within four different Time windows 73
Figure 3.3. Prediction Metrics of different Definitions of Early Improvement in Depressive Symptoms measured by EMA or WQA on Treatment Response to a seven-week Psychological Treatment 76

Study III

Figure 4.1. Data exclusion flow diagram 88
Figure 4.2. Linear Regressions of Momentary RNT Concreteness over the course of the seven-week intervention period 98

List of Tables

General Introduction

Table 1.1. Original wordings and english translations of the EMA items 26

Study I

Table 2.1. Descriptive Statistics of the Treatment Arms 44

Table 2.2. Multi-Level Models of Depression and RNT measured by EMA versus WQA among Patients randomly assigned to ST, IST or CBT, with ST as Reference Group *46*

Table 2.3. Standardized Effect Sizes (Cohen's d and NNT) of Depression and RNT measured by EMA versus WQA among patients randomly assigned to ST, IST or CBT *47*

Table 2.4. Multiple Regression Analyses predicting Global Functioning (GF) measured with Clinical Interview (WHO-DAS) after seven Weeks of Treatment based on Baseline Global Functioning and the Slope of Depression and RNT measured with EMA versus WQA 49

Table 2.5. Multi-Level Models of Depression and RNT measured by EMA versus WQA with Assessment Technique (EMA versus WQA) as Predictor 51

Study II

Table 3.1. Descriptive Statistics of the Sample 69

Table 3.2. Logistic Regression Models Predicting Treatment Response based on Early Improvement Rates measured within four different Time Windows with EMA or WQA 71

Table 3.3. Prediction Metrics of different Definitions of Early Improvement in Depressive Symptoms measured by EMA or WQA on Treatment Response to a seven-week Psychological Treatment 75

Study III

Table 4.1. Descriptive Statistics of the Sample (N = 77) 95

Table 4.2. MLM's of Momentary Depression predicted by Momentary RNT versus Momentary RNT and the Concreteness of Momentary RNT 96

List of Tables

Table 4.3. MLM of Concreteness of Momentary Repetitive Negative Thinking predicted by Time and Improvement of Depression Severity (BDI-II) from baseline to the end of the intervention 97

Table 4.4. MLM's of Momentary Depression and Momentary Concreteness predicted by the respective other variable on the previous timepoint 99

Appendix A: Supplementary Material Study I

The supplementary material provided here is intended to offer readers a more comprehensive view of the study and to ensure full transparency in the presentation of our research findings. In line with the submission guidelines, we have included six key components that are referenced in the main manuscript:

- **Table A.1:** This table provides an overview over the assessment plan of the study.
- **Table A.2:** This table provides the precise wordings of each individual EMA (Ecological Momentary Assessment) item used in our study.
- **Table A.3:** This table provides an overview over further descriptives of the sample, such as care condiction, comorbidities and co-therapies.
- **Table A.4:** This table provides the results of the comparison analyses conducted between the simple multilevel models, which included only time and intervention condition as predictors and the complex multilevel models, which were expanded to include the covariates gender, baseline depression, response rate and concomitant care.
- **Figure A.1:** This figure visually represents the outcomes of a survival analysis conducted to assess whether dropout rates were evenly distributed across all intervention groups. This analysis is an essential aspect of understanding the dropout patterns in our study and its potential impact on the overall findings.
- **Figure A.2:** This figure shows the distribution of patient response rates to the ecological momentary assessment (EMA) over the intervention weeks.

For a comprehensive understanding of the study, we encourage readers to refer to these supplementary materials when prompted in the main manuscript. We believe that they significantly contribute to the overall transparency and depth of the research presented.

Table A.1

Assessment Plan of the Study

	Baseline week							**Week 1 - 6**							**Week 7**						
Weekdays	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Intervention								2x single- &							2x single- &						
(CBT/ST/IST)								2x group-sessions							2x group-sessions						
Assessments																					
Eligibility check				X																	
EMA	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x
BDI-II/ PTQ				X							X							X			
WHODAS				X														X			

Note. This table gives a timeline of the interventions and assessments in the study. Note that only those assessments relevant for our analyses are mentioned here. CBT: Cognitive Behavioral Therapy; ST: Schema Therapy; IST: Individual Supportive Therapy; EMA: Ecological Momentary Assessment; BDI-II; Beck's Depression Inventory II; PTQ: Perseverative Thinking Questionnaire; WHO-DAS: World Health Organization Disability Assessment Schedule.

Table A.2

Wordings of the EMA Items

EMA Item construct	**Item Wording (German original)**	**Item Wording (English translation)**
Depressive symptoms		
Loss of interest	Hast du gerade das Gefühl, zu nichts mehr Lust zu haben?	Do you feel like you don't want to do anything anymore?
Withdrawal	Ziehst du dich gerade von sozialen Kontakten oder Aktivitäten zurück?	Are you currently withdrawing from social contacts or activities?
Psychomotor agitation /inhibition	Fühlst du dich gerade besonders körperlich gehemmt oder aktiviert?	Are you feeling particularly physically inhibited or agitated?
Current mood	Wie fühlst du dich?	How are you feeling?
RNT		
Repetitiveness of RNT	Dieselben negative Gedanken gehen mir immer und immer wieder durch den Kopf.	The same negative thoughts keep going through my mind again and again.
Uncontrollability of RNT	Ich hänge an bestimmten negativen Gedanken fest und kann mich nicht davon lösen.	I get stuck on certain negative issues and can’t move on.
Intrusiveness RNT	Negative Gedanken tauchen auf, ohne dass ich dies will.	Negative thoughts come to my mind without me wanting them to.
Subjective burden through RNT	Ich fühle mich durch negative Gedanken beeinträchtigt.	I feel weighted down by negative thoughts.

Note. The response scale of all EMA Items, except for the mood item, was two-stepped. Participants responded to a binary *Yes-No* scale. If *Yes* was selected, a five-point Likert scale followed, which assessed the extent of agreement (labeling: *not at all, a bit, moderately, considerably, very much*). The ‘Current mood’ item was rated by selecting one of five emojis (labeling: *very good*, *good*, *moderate*, *bad*, *very bad*).

Table A.3

Descriptive Statistics of the Treatment Arms

			Treatment								
	Total		ST		CBT		IST				
Characteristic	(N=71)		(N=20)		(N=28)		(N=23)				
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**t or Chi²**	**df**	***p***
Martital status (married or steady relationship)	31	43.66	9	45.00	15	53.57	7	30.43	2.59	2	0.27
Care condition									6.54	2	0.04
Inpatient care	46	64.79	9	45.00	18	64.29	19	82.61			
Day clinical care	25	35.21	11	55.00	10	35.71	4	17.39			
Covid-19[a]									0.92	2	0.63
Before Corona	21	29.58	5	25.00	7	25.00	9	39.13			
Partly during Corona	10	14.08	2	10.00	6	21.43	2	8.70			
During Corona	40	56.34	13	65.00	15	53.57	12	52.17			
Co-therapies											
Ergotherapy	59	83.10	18	90.00	24	85.71	17	73.91	2.17	2	0.34
Case management	62	87.32	16	80.00	25	89.29	21	91.30	1.38	2	0.50
Relaxation training	26	36.62	6	30.00	11	39.29	9	39.13	0.52	2	0.77
Cognitive training	2	2.82	1	5.00	1	3.57	0	0.00	1.06	2	0.59
Sports	35	49.30	6	30.00	15	53.57	14	60.87	4.35	2	0.11
Axis I comorbidities[b]											
Agoraphobia	3	4.23	2	11.11	0	0.00	1	4.76	0.19	2	0.91
Panic disorder	7	9.86	2	11.11	4	14.81	1	4.76	1.81	2	0.40
Panic disorder with agoraphobia	8	11.27	1	5.56	3	11.11	4	19.05	1.32	2	0.52
Social phobia	15	21.13	4	22.22	5	18.52	6	28.57	0.23	2	0.89
Specific phobia	15	21.13	5	27.78	5	18.52	5	23.81	0.02	2	0.99
Generalized anxiety disorder	9	12.68	2	11.11	2	7.41	5	23.81	1.08	2	0.58
Post-traumatic stress disorder	12	16.90	5	27.78	2	7.41	5	23.81	0.97	2	0.62
Obsessive compulsive disorder	6	8.45	1	5.56	2	7.41	3	14.29	0.73	2	0.69
Eating Disorder	3	4.23	1	5.56	0	0.00	2	9.52	0.25	2	0.88
Substance disorder	13	18.31	5	27.78	2	7.41	6	28.57	1.37	2	0.50
Somatoform disorder	19	26.76	8	44.44	3	11.11	8	38.10	3.10	2	0.21

	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	**t or Chi²**	**df**	***p***
Medication (no. of weeks)											
Antidepressants	6.56	2.69	6.20	3.19	6.04	3.01	7.52	1.31	4.22	2	0.12
Neuroleptics	1.83	3.18	0.80	2.46	2	3.27	2.52	3.51	4.12	2	0.13
Tranquilizer	0.80	2.09	0.90	2.27	1.18	2.45	0.26	1.25	2.95	2	0.23
Mood Stabilizer	0.34	1.44	0.05	0.22	0.54	1.73	0.35	1.67	0.96	2	0.62

Note. ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; BDI-II; Beck's Depression Inventory II; PTQ: Perseverative Thinking Questionnaire; WHO-DAS: World Health Organization Disability Assessment Schedule.

[a]The start of the Covid-19 pandemic was set at 10th March 2020, coinciding with the enforcement of initial hygiene measures and mandatory visiting restrictions for staff and patients in our clinic.

[b]Comorbidities were diagnosed by the Munich-Composite International Diagnostic Interview (M-CIDI; Wittchen et al., 1998), which is a computerized, fully standardized German version of the World Mental Health Composite Internatinal Diagnostic Interview (WHO-CIDI; Robins et al., 1988).

Table A.4

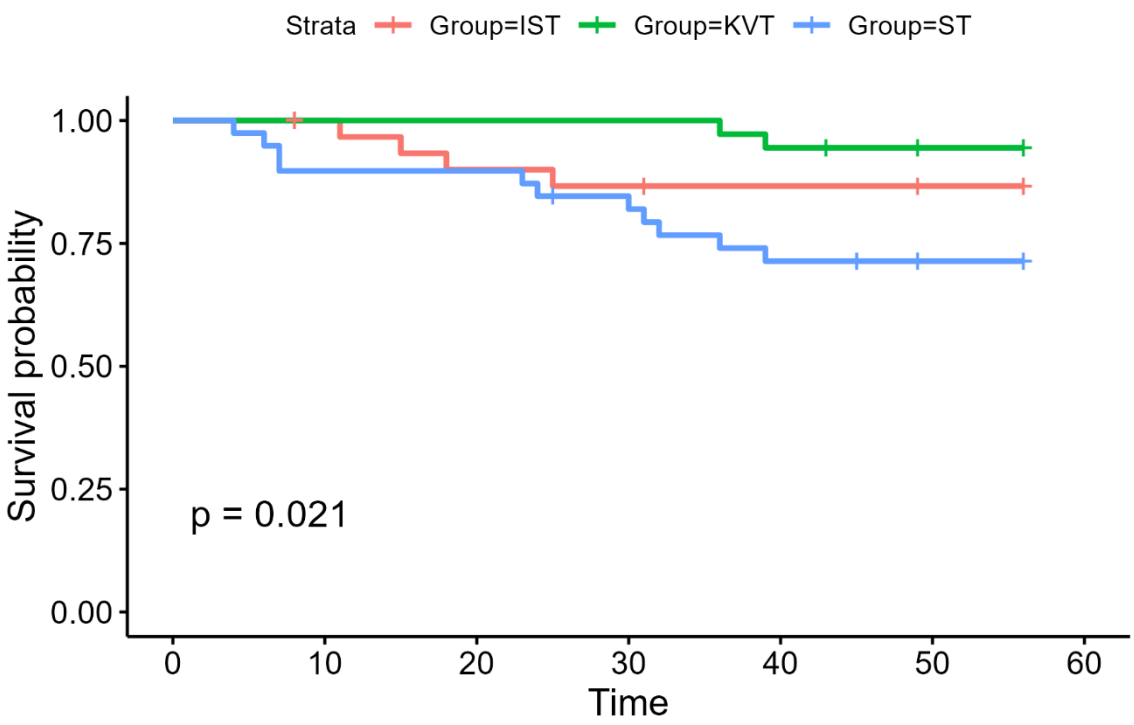
Comparison Analyses between simple and complex Multilevel Models

Model	**df**	**AIC**	**BIC**
EMA – Depressive symptoms			
Simple MLM	10	19038.21	19106.59
Complex MLM	27	19067.14	19251.76
WQA – Depressive symptoms			
Simple MLM	10	1119.69	1162.111
Complex MLM	27	1150.25	1264.794
EMA – RNT			
Simple MLM	10	19109.75	19178.13
Complex MLM	27	19138.50	19323.13
WQA – RNT			
Simple MLM	10	1276.71	1319.14
Complex MLM	27	1307.45	1421.99

Note. This table provides the results of the comparison analyses conducted between the simple multilevel models, which included only time and intervention condition as predictors and the complex multilevel models in those the covariates gender, baseline depression, response rate and concomitant care were added. RNT: Repetitive Negative Thinking; EMA: Ecological Momentary Assessment; WQA: Weekly Questionnaire Assessment; MLM: Multi Level Model.

Figure A.1

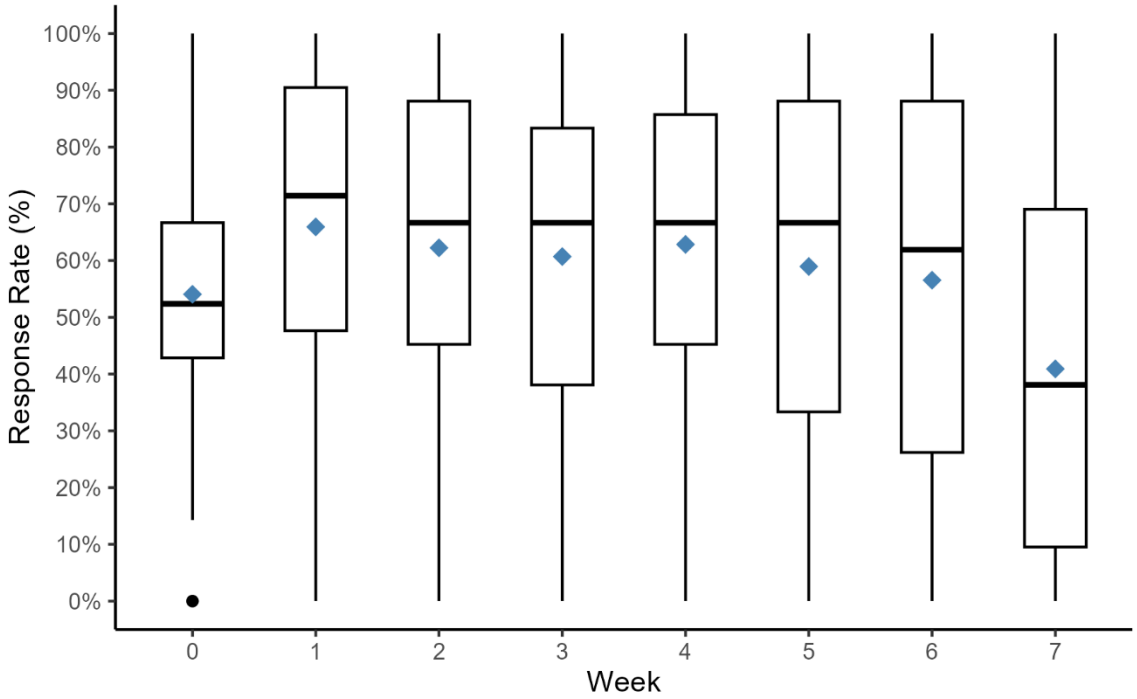
Proportion of Patients in the Treatment Arms (ST, IST, or CBT) throughout the Baseline Week and the seven-week Treatment Phase



Note. N = 106 (Intent-to-treat sample); ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; Drop-outs during the study were defined as enrolled patients who left the clinic before end of intervention, who missed more than six sessions of their intervention (equivalent to 22% of the total intervention dose) or requested for a different treatment.

Figure A.2

Distribution of Patient Response rates to the Ecological Momentary Assessment over the intervention weeks



Note. N = 71; Week: Intervention week; Comparison of patient response rates to the ecological momentary assessment over the intervention weeks. The boxplots show the distribution of the response rates for each treatment week. The box represents the interquartile range (IQR), with the line inside representing the median and the blue square representing the mean response rate. The whiskers extend to the minimum and maximum values within 1.5 times the IQR above the quartiles. Outliers are represented by individual points outside the whiskers.

Appendix B: Supplementary Material Study II

The supplementary material provided here is intended to offer readers a more comprehensive view of the study and to ensure full transparency in the presentation of our research findings. In line with the submission guidelines, we have included six key components that are referenced in the main manuscript:

- **Table B.1:** This table provides an overview over the assessment plan of the study.
- **Table B.2:** This table provides the precise wordings of each individual EMA (Ecological Momentary Assessment) item used in our study.
- **Table B.3:** This table provides an overview over further descriptives of the sample, such as care condiction, comorbidities and co-therapies.
- **Table B.4:** This table provides on overview of the four ROC-analyses and the comparison analyses of the AUC values.
- **Figure B.1:** This figure shows the distribution of patient response rates to the ecological momentary assessment (EMA) over the intervention weeks.
- **Figure B.2:** This figure shows the improvement rate distributions of the four time windows separately for responders and non-responders and the two variables EMA and WQA.

For a comprehensive understanding of the study, we encourage readers to refer to these supplementary materials when prompted in the main manuscript. We believe that they significantly contribute to the overall transparency and depth of the research presented.

Table B.1

Assessment Plan of the Study

	Baseline week							**Week 1 - 4**							**Week 7**						
Weekdays	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Intervention								2x single- &							2x single- &						
(CBT/ST/IST)								2x group-sessions							2x group-sessions						
Assessments																					
Eligibility check				X																	
EMA	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x
BDI-II				X							X							X			

Note. This table gives a timeline of the interventions and assessments in the study. Note that only those assessments relevant for our analyses are mentioned here. CBT: Cognitive Behavioral Therapy; ST: Schema Therapy; IST: Individual Supportive Therapy; EMA: Ecological Momentary Assessment; BDI-II; Beck's Depression Inventory II.

Table B.2

Wordings of the EMA Items

EMA Item construct	**Item Wording (German original)**	**Item Wording (English translation)**
Loss of interest	Hast du gerade das Gefühl, zu nichts mehr Lust zu haben?	Do you feel like you don't want to do anything anymore?
Withdrawal	Ziehst du dich gerade von sozialen Kontakten oder Aktivitäten zurück?	Are you currently withdrawing from social contacts or activities?
Psychomotor agitation /inhibition	Fühlst du dich gerade besonders körperlich gehemmt oder aktiviert?	Are you feeling particularly physically inhibited or agitated?
Current mood	Wie fühlst du dich?	How are you feeling?

Note. The response scale of all EMA Items, except for the mood item, was two-stepped. Participants responded to a binary *Yes-No* scale. If *Yes* was selected, a five-point Likert scale followed, which assessed the extent of agreement (labeling: *not at all, a bit, moderately, considerably, very much*). The 'Current mood' item was rated by selecting one of five emojis (labeling: *very good*, *good*, *moderate*, *bad*, *very bad*).

Table B.3

Descriptive Statistics of the Sample

Characteristic	Total (N=52)		Responder Status: Responder (N=25)		Responder Status: Non-Responder (N=27)				
	N	**%**	**N**	**%**	**N**	**%**	**Chi²**	**df**	***p***
Intervention condition							1.57	2	0.455
CBT	21	40.38	10	40.00	11	40.74			
ST	15	28.85	9	36.00	6	22.22			
IST	16	30.77	6	24.00	10	37.04			
Gender (female)	30	57.69	15	60.00	15	55.56	<0.00	1	0.966
Nationality (german)	44	84.62	22	88.00	22	81.48	1.07	2	0.586
Martital status									
(married or steady relationship)	25	48.08	15	60.00	10	37.04	3.31	2	0.191
School graduation									
(Qualification for University entrance)	27	51.92	14	56.00	13	48.15	2.96	2	0.227
Income							6.21	3	0.102
Low income (up to 1500 EUR)	20	38.46	8	32.00	12	44.44			
Middle income (1500 - 4000 EUR)	19	36.54	11	44.00	8	29.63			
High income (more than 4000 EUR)	9	17.31	6	24.00	3	11.11			
not specified	4	7.69	0	0.00	0	0.00			
Care condition							<0.00	1	1.000
Inpatient care	31	59.62	15	60.00	16	59.26			
Day clinical care	21	40.38	10	40.00	11	40.74			
Covid-19[a]							0.80	2	0.670
Treated before the pandemic	15	28.85	7	28.00	8	29.63			
Treated partly during the pandemic	8	15.38	5	20.00	3	11.11			
Treated during the pandemic	29	55.77	13	52.00	16	59.26			
Co-therapies									
Ergotherapy	46	88.46	19	76.00	27	100.00	5.16	1	0.023
Case management	46	88.46	20	80.00	26	96.30	1.97	1	0.161
Relaxation training	22	42.31	5	20.00	17	62.96	8.14	1	0.004
Cognitive training	2	3.85	2	8.00	0	0.00	0.60	1	0.437
Sports	26	50.00	13	52.00	13	48.15	<0.00	1	1.000
Comorbidities[b]									
Agoraphobia	1	1.92	0	0.00	1	4.17	1.95	2	0.378
Panic disorder	4	7.69	2	8.33	2	8.33	0.92	2	0.630
Panic disorder with agoraphobia	9	17.31	4	16.67	5	20.83	1.06	2	0.588
Social phobia	9	17.31	3	12.50	6	25.00	2.16	2	0.340
Specific phobia	11	21.15	6	25.00	5	20.83	1.04	2	0.594

Generalized anxiety disorder	5	9.62	3	12.50	2	8.33	1.15	2	0.563
Post-traumatic stress disorder	11	21.15	6	25.00	5	20.83	1.04	2	0.594
Obsessive compulsive disorder	3	5.77	0	0.00	3	12.50	4.13	2	0.127
Eating Disorder	2	3.85	1	4.17	1	4.17	0.92	2	0.630
Substance disorder	8	15.38	4	16.67	4	16.67	0.92	2	0.630
Somatoform disorder	13	25.00	6	25.00	7	29.17	1.03	2	0.597
	M	***SD***	***M***	***SD***	***M***	***SD***	**t or W**	**df**	***p***
Age (years)	40.5	11.71	43.12	12.04	38.07	11.06	-1.57	48.70	0.123
Response Rates									
EMA	64.67	23.35	67.84	20.44	61.73	25.80	-0.95	48.87	0.347
BDI	99.04	3.36	99	3.46	99.07	3.34	339.50		0.953
Baseline Symptoms									
BDI	32.58	8.50	32.84	8.42	32.33	8.73	-0.21	49.91	0.832
Medication (number of weeks)[c]									
Antidepressants	6.63	2.63	7.44	1.50	5.89	3.20	261.50		0.097
Neuroleptics	1.63	3.09	1.96	3.32	1.33	2.88	293.50		0.293
Tranquilizer	0.87	2.25	0.64	2.22	1.07	2.30	368.00		0.354
Mood Stabilizer	0.46	1.67	0.64	1.82	0.30	1.54	297.50		0.157

Note. ST: Schema Therapy; CBT: Cognitive Behavioral Therapy; IST: Individual Supportive Therapy; BDI-II: Beck's Depression Inventory II; EMA: Ecological Momentary Assessment.

[a]The start of the Covid-19 pandemic was set at 10th March 2020, coinciding with the enforcement of initial hygiene measures and mandatory visiting restrictions for staff and patients in our clinic.

[b]Comorbidities were diagnosed by the Munich-Composite International Diagnostic Interview (M-CIDI; Wittchen et al., 1998), which is a computerized, fully standardized German version of the World Mental Health Composite Internatinal Diagnostic Interview (WHO-CIDI; Robins et al., 1988).

[c]For medications the number of weeks a certain medication was given is reported.

Table B.4

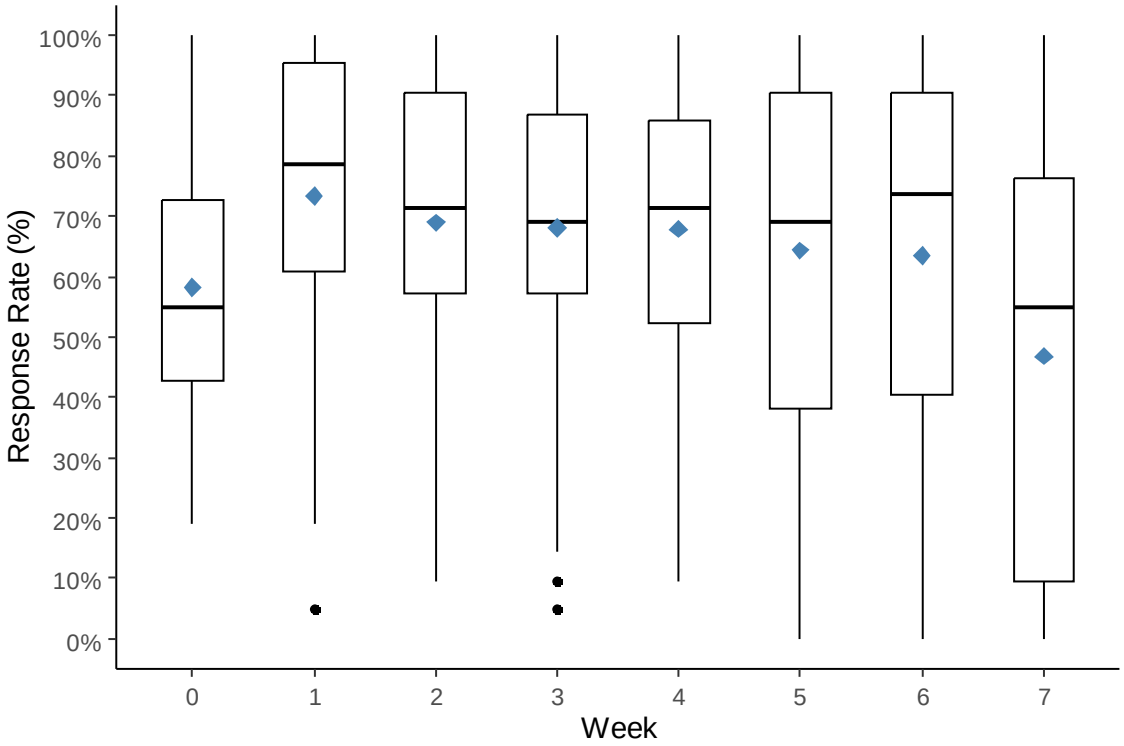
ROC-Analyses and Comparisons of the AUC Values for Predicting Treatment Response Based on Early Improvement Rates Measured with EMA and WQA

		Maximum Youden Index		AUC model comparison analysis (De Longs Test)		
Time window	**AUC**	**Y**	**Early Improvement rate**	**Comparing time window**	**z (*p*)**	**95% CI**
EMA						
time window 1	0.57	0.22	16.82%	time window 2	-1.28 (0.199)	[-0.28 - 0.06]
time window 2	0.68	0.42	17.08%	time window 3	-0.09 (0.932)	[-0.14 - 0.13]
time window 3	0.68	0.41	29.16%	time window 4	-1.25 (0.21)	[-0.12 - 0.03]
time window 4	0.73	0.41	24.7%			
WQA						
time window 1	0.65	0.32	24.74%	time window 2	-0.33 (0.744)	[-0.19 - 0.13]
time window 2	0.67	0.4	11.1%	time window 3	-0.51 (0.61)	[-0.13 - 0.08]
time window 3	0.7	0.35	21.23%	time window 4	-1.47 (0.143)	[-0.17 - 0.02]
time window 4	0.77	0.45	42.64%			
EMA versus WQA						
EMA time window 1				WQA time window 1	-0.72 (0.471)	[-0.29 – 0.14]
EMA time window 2				WQA time window 2	0.07 (0.947)	[-0.17 – 0.18]
EMA time window 3				WQA time window 3	-0.17 (0.863)	[-0.18 – 0.15]
EMA time window 4				WQA time window 4	-0.54 (0.59)	[-0.19 – 0.11]

Note. EMA: Ecological Momentary Assessment; WQA: Weekly questionnaire assessment; AUC: Area Under the Curve; $n = 52$. This table presents Receiver Operating Characteristic (ROC) models predicting treatment response (vs. non-reponse) through early improvement rates in depressive symptoms measured with EMA or WQA. Early improvement rates were measured within four time windows, namely one, two, three or four weeks after treatment initiation. The AUC values represent how well the models discriminate between responders and non-responders - higher AUC values indicate better performance. Additionally, for each time window the early improvement rate with the highest youden index (Y) is denoted. Finally, the AUC values of the different ROC models were compared with model comparsion analyses (De Longs Test), indicating no significant differences between the compared models.

Figure B.1

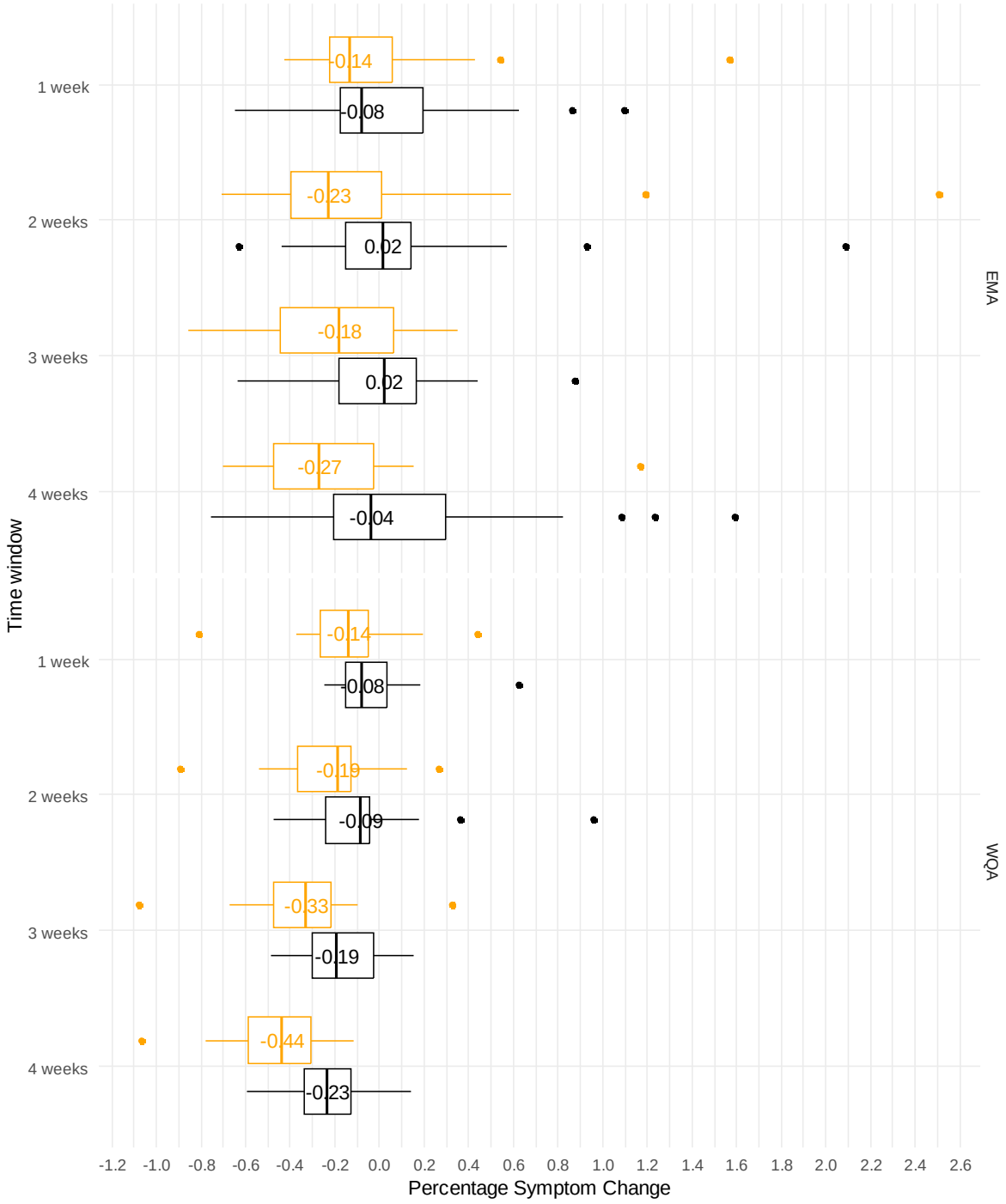
Distribution of Patient Response rates to the Ecological Momentary Assessment over the intervention weeks



Note. N = 52; Week: Intervention week; Comparison of patient response rates to the ecological momentary assessment over all seven intervention weeks. The boxplots show the distribution of the response rates for each treatment week. The box represents the interquartile range (IQR), with the line inside representing the median and the blue square representing the mean response rate. The whiskers extend to the minimum and maximum values within 1.5 times the IQR above the quartiles. Outliers are represented by individual points outside the whiskers.

Figure B.2

Improvement Rate Distributions of the four Time Windows separately for Responders and Non-Responders and the two variables EMA and WQA



Note. Dep: Depression; EMA: Ecological Momentary Assessment; WQA: Weekly Questionnaire Assessment; Sample size in all models is 52.

Appendix C: Supplementary Material Study III

The supplementary material provided here is intended to offer readers a more comprehensive view of the study and to ensure full transparency in the presentation of our research findings. We have included six key components that are referenced in the main manuscript:

- **Table C.1:** This table provides an overview of the assessment plan of the study.
- **Table C.2:** This table provides an overview of the exact wording of the EMA-items as well as an English translation.
- **Table C.3:** This table provides the results of the MLM of concreteness of momentary RNT predicted by time and improvement of depression severity (BDI-II) with control for therapy groups from baseline to the end of the intervention.
- **Table C.4:** This table provides the results of the comparison analyses conducted between the simple multilevel models and the complex multilevel models, which included the covariates age and gender.
- **Figure C.1:** This figure shows the distribution of the frequency of text units and the number of words written.
- **Figure C.2:** This figure demonstrates the decision tree developed to rate text units for concreteness.

For a comprehensive understanding of the study, we encourage readers to refer to these supplementary materials when prompted in the main manuscript. We believe that they significantly contribute to the overall transparency and depth of the research presented.

Table C.1

Weekly assessments of EMA and BDI-II measures of the OPTIMA study

	Baseline week							**Week 1 - 7**						
Week days	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Treatment										2x single- &				
(CBT/ST/IST)										2x group-sessions				
Assessments														
BDI-II				X							X			
EMA	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x	3x

Note. ST: Schema Therapy; CBT: Cognitive Behavioural Therapy; IST: Individual Supportive Therapy.

Table C.2

Wordings of the EMA Items

EMA Item construct	**Item Wording (German original)**	**Item Wording (English translation)**
Depressive symptoms		
Loss of interest	Hast du gerade das Gefühl, zu nichts mehr Lust zu haben?	Do you feel like you don't want to do anything anymore?
Withdrawal	Ziehst du dich gerade von sozialen Kontakten oder Aktivitäten zurück?	Are you currently withdrawing from social contacts or activities?
Psychomotor agitation /inhibition	Fühlst du dich gerade besonders körperlich gehemmt oder aktiviert?	Are you feeling particularly physically inhibited or agitated?
Current mood	Wie fühlst du dich?	How are you feeling?
RNT		
Repetitiveness of RNT	Dieselben negative Gedanken gehen mir immer und immer wieder durch den Kopf.	The same negative thoughts keep going through my mind again and again.
Uncontrollability of RNT	Ich hänge an bestimmten negativen Gedanken fest und kann mich nicht davon lösen.	I get stuck on certain negative issues and can’t move on.
Intrusiveness RNT	Negative Gedanken tauchen auf, ohne dass ich dies will.	Negative thoughts come to my mind without me wanting them to.
Subjective burden through RNT	Ich fühle mich durch negative Gedanken beeinträchtigt.	I feel weighted down by negative thoughts.
Free-text RNT-item (which was later rated for concreteness)	Welche negativen Gedanken gehen dir aktuell wiederholt durch den Kopf? Bitte schreibe deine Gedanken in ganzen Sätzen auf.	Which negative thoughts are currently going through your mind repeatedly? Please write down your thoughts in complete sentences"

Note. Except for the mood- and the free-text RNT-item, all EMA items utilized a two-step response scale. Participants initially answered a binary Yes-No question. If 'Yes' was chosen, they were then prompted to use a five-point Likert scale to indicate the degree of agreement, with labels ranging from 'not at all' to 'very much'. For the 'Current mood' item, participants selected from five emojis representing different levels of mood, with labels including 'very good', 'good', 'moderate', 'bad', and 'very bad'.

Table C.3

MLM of Concreteness of Momentary Repetitive Negative Thinking predicted by Time and Improvement of Depression Severity (BDI-II) with control for therapy groups from baseline to the end of the intervention

IDV/predictors	**Estimates**	**SE**	**t**	***p***	**95% CI**
Intercept	2.622123	0.22	11.70	<.001	[2.174603 – 3.064105]
Time	-0.003226	0.00	-4.03	<.001	[-0.004795– -0.001657]
BDI-II Impr	-0.002303	0.00	-0.55	.585	[-0.010598 – 0.006050]
Group [1]	-0.476930	0.37	-1.30	.199	[-1.204295 – 0.257048]
Group [2]	-0.623270	0.39	-1.61	.112	[-1.389897 – 0.150254]
BDI-II Impr*Time	0.000049	0.00	3.41	<.001	[0.000021 – 0.000077]
BDI-II Impr*Group [1]	0.006647	0.01	0.99	.327	[-0.006757 – 0.019970]
BDI-II Impr*Group [2]	0.008725	0.01	1.15	.253	[-0.006348 – 0.023714]
Time*Group [1]	0.001352	0.00	1.03	.304	[-0.001229 – 0.003933]
Time*Group [2]	-0.000351	0.00	-0.25	.806	[-0.003155 – 0.002453]
BDI-II Impr*Time*Group [1]	-0.000008	0.00	-0.33	. 743	[-0.000056 – 0.000040]
BDI-II Impr*Time*Group [2]	0.000435	0.00	1.44	.149	[-0.000016 – 0.000103]

Note. n = 77; ICC = 0.34; Marginal R^2 = 0.038, Conditional R^2 = 0.365; Estimates = unstandardised regression coefficients; BDI-II: Beck's Depression Inventory; BDI-II Impr = BDI-II Improvement; Group was dummy coded with 0 = cognitive behavioural therapy, 1 = schema therapy and 2 = individual supportive therapy.

Table C.4

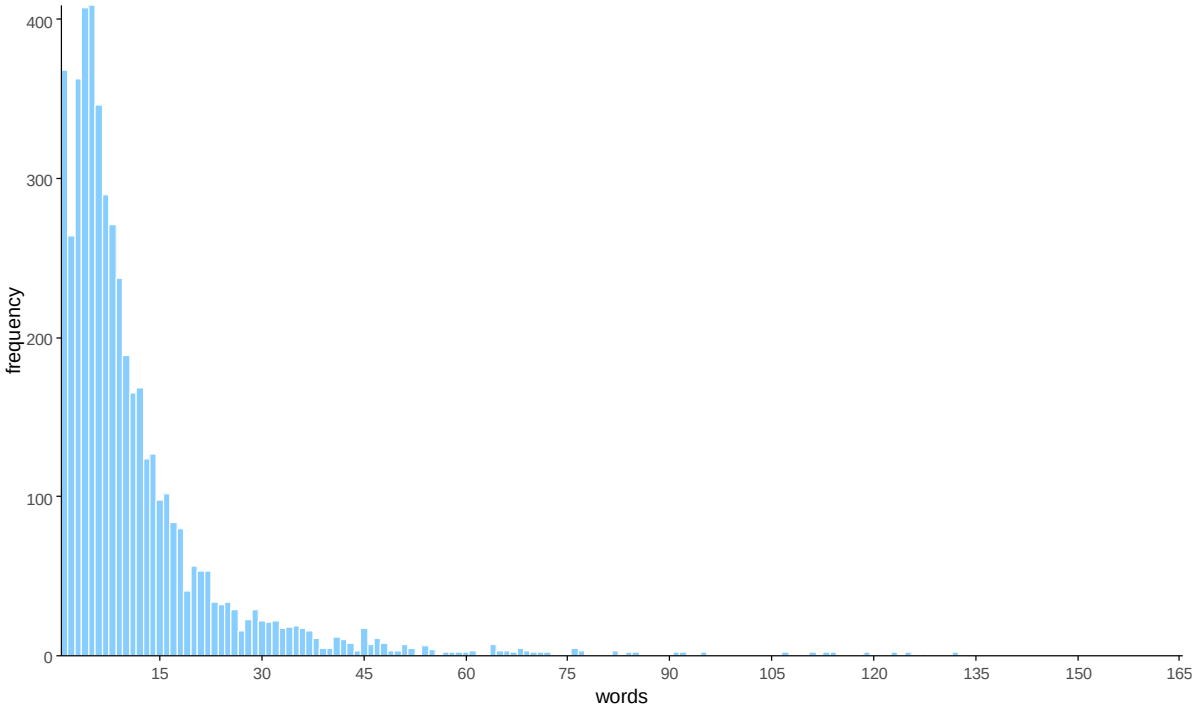
Comparison Analyses between simple and complex Multilevel Models

Model	df	AIC	BIC
Prediction of momentary depression by the concreteness of momentary RNT			
Momentary Depression (Model A)			
Simple MLM		9496.8	9535.8
Complex MLM	2	9500.0	9552.0
Momentary Depression (Model B)			
Simple MLM		9407.8	9459.7
Complex MLM	2	9410.9	9475.8
Prediction of Concreteness of RNT by the time and BDI-II improvement			
Simple MLM		11498	11538
Complex MLM	2	11501	11553
Temporal relationship between momentary concreteness and momentary depression			
Momentary Depression			
Simple MLM		7173.9	7204.4
Complex MLM	2	7176.0	7218.7
Concreteness of MomRNT			
Simple MLM		7709.4	7739.9
Complex MLM	2	7711.6	7754.3

Note. This table provides the results of the comparison analyses conducted between the simple multilevel models, which included the predictors described in the statistical modelling section and the complex multilevel models including the control variables age and gender. RNT: Repetitive Negative Thinking; BDI-II: Beck's Depression Inventory; MLM: Multi Level Model.

Figure C.1

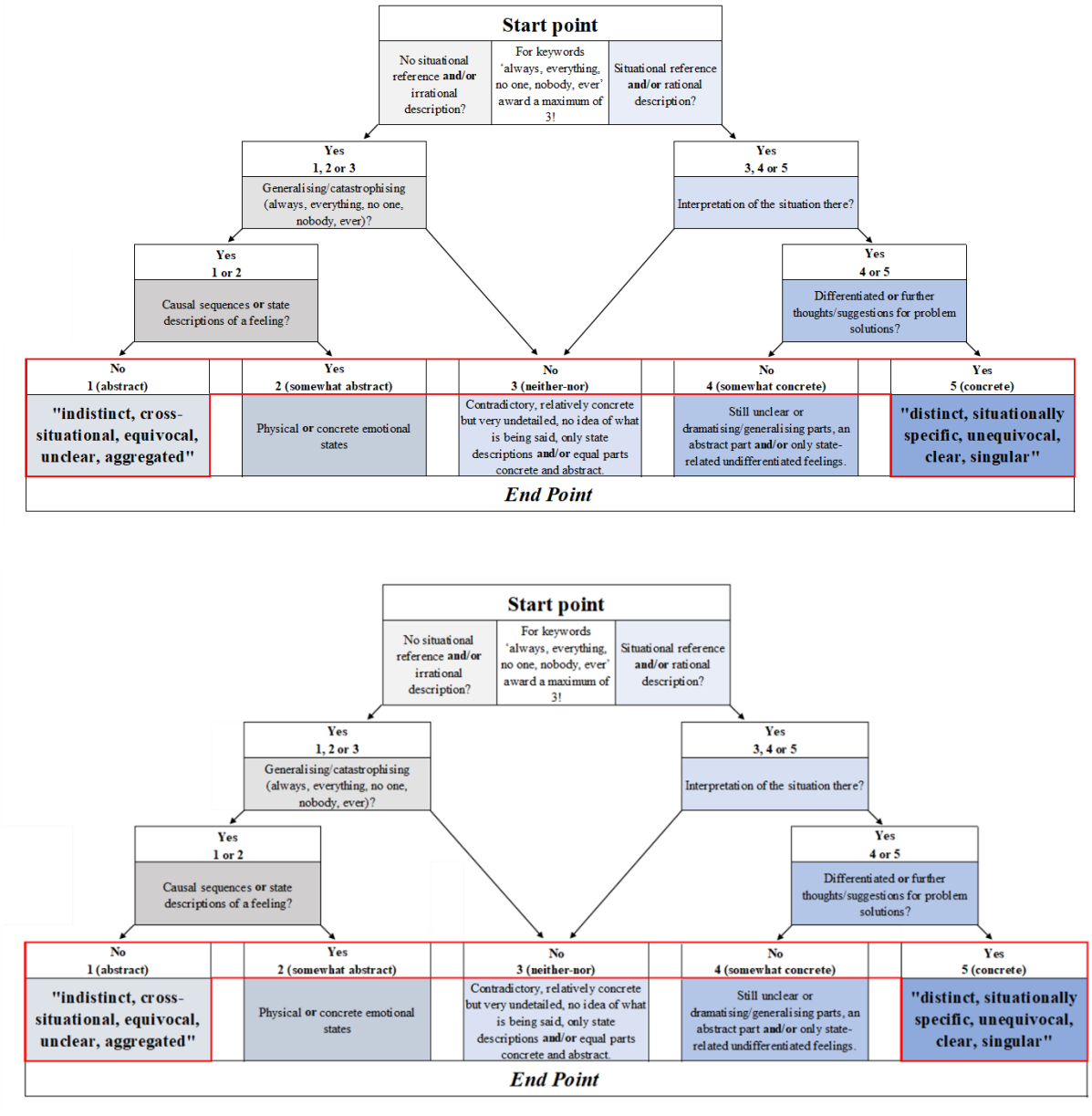
Distribution of the frequency of text units and their number of words written



Note. The graph presents how many words were written and how frequent this appeared in the sample.

Figure C.2

Decision aid (german original and english translation)



Note. The german version of the decision tree, developed based on the application of Stöber's (2002) definitions of concreteness and the concreteness-scale (presented in the red frame) on the present sample.