Patient outcomes from open-ended psychotherapy in a real-world setting:

An investigation of psychotherapeutic change

Magnus Nordmo

Department of Psychology

Faculty of Social Science

University of Oslo

2020

Series of dissertations submitted to the
Faculty of Social Sciences, University of Oslo
No. 832

ISSN 1564-3991

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgments

The results presented in this thesis derives from decades of work by clinicians, patients and other researchers. I am truly grateful to be given the opportunity to analyze a dataset that has required countless hours to produce. Skimming the proverbial cream of the Norwegian Multisite Study of Process and Outcome in Psychotherapy project has been rewarding both as a researcher and a clinician. I would like to particularly acknowledge the patients who have invested time and effort in their participation in the project. This could not have been done without your efforts. This statement is also true for the clinicians and site managers who have diligently prioritized research in a hectic clinical practice.

I want to express my deep gratitude for the guidance provided by my two supervisors, Ole Andre Solbakken and Jon Trygve Monsen. Your dedication to values promoting patient recovery and well-being has been a great motivator and inspiration for this work. You have also thought me more than a fair share of practical research methodology, statistics and clinical theory that I will take with me throughout my career. Ole Andre, I have always felt welcome into your office even though I know you have a very hectic schedule. You have always encouraged my ideas which have given me a personal sense of academic self-efficacy. This process has been a critical impetus in the face of advanced and challenging topics.

Another source of academic inspiration has come from countless enjoyable hours of engaging but lighthearted conversations with fellow Ph.D. candidates and friends. Our discussions and musings over everything and nothing have helped me to appreciate the perplexity of psychology and has served as a great opportunity for unfiltered conversations regarding the realities of doing science. Thank you Nikolai Haahjem Eftedal, Erik Nakkerud and Thomas Haarklau Kleppestø and Morten Nordmo.

Lastly, I am indebted to my dear wife, Lene Caroline Ljosland Nordmo, who has been a continuous source of encouragement and support during this period. You are my greatest ally in the face of challenging deadlines and stressful life events. I could not have done this without you.

Summary

There has been substantial development and expansion of the knowledge base in psychotherapy research in recent years. The effectiveness of psychological treatment compared to wait-list control, the positive relationship between outcomes and alliance, the differential effectiveness of therapists, the equivalence of outcomes across different psychotherapy models, the superiority of focused as opposed to unstructured interventions, and the positive effect of patient monitoring have all been empirically supported through a voluminous research literature. However, the majority of studies demonstrating these findings have been on patients with, relatively speaking, lower levels of disorder complexity and psychological dysfunction, often excluding those with multiple comorbid disorders and severe character-based pathology. Treatments are often short and manualized. Thus, there is a notable lack of studies demonstrating the effectiveness of common treatment approaches with more complex patient-populations. Since the majority of patients in many specialist psychotherapy services can be classified as complex cases with multiple disorders and commonly have personality-based pathology in addition to their symptom disorders, this is a very unfortunate limitation in the existing literature.

The Norwegian Multisite Study of Process and Outcome in Psychotherapy – (NMSPOP) is designed to remedy these limitations. It includes naturalistically selected patients from clinical specialist services, half of which satisfy criteria for at least one personality disorder at treatment onset. The selection procedure was designed to incorporate a highly heterogeneous sample ranging from patients with highly complex and severe disorders to patients presenting mild and limited psychopathology. Furthermore, the NMSPOP is designed so that the majority of treatments delivered are in the open-ended format of psychotherapy. This entails that therapists and patients are instructed to come to an agreement about when to terminate treatment based on the patient's difficulties and progress or lack

thereof. Outcomes are measured in terms of both self-rated questionnaires but also in terms of observer-rated diagnostic changes. This dual approach of outcome evaluation adds to the overall measurement validity as both the patient and a trained observer serve as evaluators of progress.

In paper I, we explore the main outcomes of the project. Our main finding was that patients experienced large positive changes in self-report measures of overall psychiatric symptoms and moderate positive changes in self-reported interpersonal problems. Improvements were stable throughout a two-and-a-half-year follow-up period. Improvements were also mirrored with observer-rated diagnostic changes, which revealed that a substantial majority recovered from their respective Axis I (58 %) and/or Axis II (55 %) disorders during treatment. The diagnostic changes were also shown to be stable throughout the follow-up period. In contrast, self-reported occupational functioning showed minimal improvement throughout the treatment and follow-up phase. Using growth-curve predicted scores, we found a surprisingly low amount of reliable patient deterioration (1 - 3 %) compared to what is generally found in the adult psychotherapy literature (5 - 10 %).

In paper II, we analyze and contrast overall outcomes in patients with and without a personality disorder at pretreatment. The results revealed that patients demonstrated equal symptomatic improvement, regardless of personality disorder status before treatment. However, patients with a personality disorder showed greater interpersonal improvement compared to those without. Observer-rated diagnostic changes were equivalent across the two groups. Similarly, both groups demonstrated equivalent and enduring improvements when assessed at a two-and-half-year follow-up. Furthermore, the degree of personality pathology was positively related to the magnitude of self-reported symptomatic improvements, as well as improvements in interpersonal functioning. Thus, patients with more severe personality problems demonstrated greater gains in the open-ended treatment format. We argue that our

results highlight the need for flexible treatment alternatives for patients with complex and characterologically-based pathology.

In paper III we explore the rate and magnitude of change during and after psychotherapy. Our results indicated that the degree of symptomatic improvement was linearly associated with time spent in psychotherapy and contingent upon the severity of psychological problems at intake. The least severely afflicted received the shortest treatments, experienced the most rapid change, but demonstrated smaller overall magnitudes of improvement. More severely suffering patients received longer treatments, had slower rates of change, but received greater overall benefits. We conclude that the rate of improvement for psychiatric symptoms and interpersonal problems during psychotherapy varies greatly between patients and that this variation is closely linked to the severity of psychopathology. Our estimates of appropriate psychotherapy dosage are much greater than what is found in the dominant effectiveness outcome literature, which is mainly focused on estimates from university counseling centers.

In summary, this work adds to an increasing knowledge base regarding the effectiveness of psychotherapy as it is delivered in a naturalistic, real-world setting. It demonstrated that most patients from a population with highly varying problem severity and disorder complexity, attain significant and stable improvements when afforded the open-ended treatment format. Another key finding is that this treatment format is associated with larger gains for patients with severe conditions than compared to patients with less severe psychopathology. These findings have implications for research and clinical practice. Firstly, patients suffering from moderate to severe psychopathology will likely benefit from a flexible treatment format. From a cost-benefit perspective, this entails that treatment lasts longer but is also associated with larger benefits. Second, when assessed during a flexible treatment, patient improvements appear to gradually emerge throughout treatment. This is in contrast to

the common claim that the majority of the benefit from psychotherapy is to be expected early in treatment, with longer treatment providing little extra benefit. Thirdly, even with a flexible and open-ended treatment, a minority of patients do not respond to treatment and some see their condition worsen. This observation may serve as an impetus for monitoring of patient outcomes and intervening if patients fail to respond to treatment or demonstrate deterioration.

List of papers

Paper I:

Effectiveness of Open-Ended Psychotherapy under Clinically Representative Conditions

Paper II:

Comparing the magnitude of improvement for patients with and without a personality disorder during open-ended psychotherapy

Paper III:

Problem severity, treatment duration, and the outcome of psychotherapy: The benefits keep growing with time spent in treatment for longer than previously known

Table of Contents

1. Introduction

Modern psychotherapy practice can trace its roots back to the work of Sigmund Freud (1856 – 1939) and the development of the psychoanalytic method. Freud discovered that patients with conditions that were previously thought to arise from purely biological processes could be treated by "a talking cure" (Breuer & Freud, 2010). The clinical effect of psychoanalytic treatment was publicly documented as case presentations, as was the norm for medical research at the time (Nissen & Wynn, 2014). The interest in measuring the changes seen as patients underwent psychotherapy, using the statistical approach we recognize today, emerged several decades later with the advent of evidence-based medicine and clinical trials. Early efforts to document the efficacy of psychotherapy treatments, such as the efforts of the humanistic school (Rogers, 1951), were characterized by small samples, unstandardized measures of pathology, and loosely defined treatments. Researchers were mainly interested in outcomes as a tool to further develop theory and develop treatment techniques. The idea of a measuring change in contrast to a control group originated with the work of Rosenthal & Frank (1956) and tied psychotherapy outcome research together with a modern medical methodology that was firmly established by 1960. Today, research on the process and outcomes of psychotherapy has come a long way and now represents a critical factor in the development of clinical expertise as well as informing policymakers. This fact is reflected by the American Psychological Association's (APA) policy statement (2006) on evidence-based practice (EBP) which includes "the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences" (p. 272). Decades of research have documented the efficacy of psychotherapy across a range of mental illnesses and has examined the conditions that influence its effects. Examples include how patient characteristics and different treatments influence overall outcomes (Gonzalez, 2016; Lambert, 2017), how alternative methods of delivery can be utilized in disseminating treatment

(Andersson et al., 2019; Firth et al., 2018; Imel et al., 2017; Nielsen et al., 2017; van Gelderen et al., 2018), and the exploration of why some therapists perform better than others (Stiles & Horvath, 2017). The influence of psychotherapy research is also evident from clinical practice guidelines that permeate clinical practice around the world. Prominent examples include the APA professional guidelines (American Psychological Association, 2015) and the National Institute for Health and Care Excellence (NICE) guidelines. The latter are ubiquitous in the United Kingdom as it laid the foundation for the Improving Access to Psychological Therapies (IAPT) program (Clark, 2011). Also worth mentioning are the many nationally coordinated psychotherapy guidelines found throughout western countries (Parry et al., 2003).

1.1 Does Therapy Work?

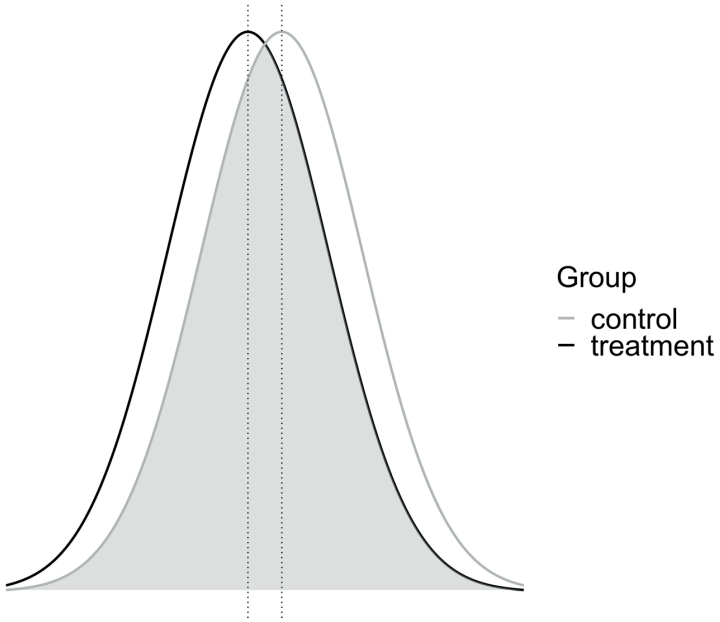
It is generally accepted that psychotherapy can produce meaningful and lasting positive change for individuals suffering from mental illness (Wampold, 2019). Although this statement is fairly uncontroversial today, it has not always been this way. The early years of psychotherapy research were fraught with controversies and intense debate. Eysenck (1952) famously pointed out that patients with a mental illness were in a state of *neurotic instability* which might make the therapist falsely attribute random fluctuations as effects of treatment. His empirical investigation concluded that patients' rates of improvement did not exceed spontaneous remission of symptoms (Eysenck, 1952, 1966). He found that the longer a patient engaged in therapy, the less likely they were to attain recovery. It should be noted that this damming conclusion was directed at long-term psychodynamic and humanistic therapy. Eysenck was an advocate of behavioral therapy which he believed was scientifically valid and produced better treatment outcomes.

Eysenck's conclusions were quickly followed by counterarguments and prolonged debate. Rosenzweig (1954) and Strupp (1963) spearheaded the effort with new analyses and interpretations which supported the notion that non-behavioral therapy can produce added

benefit over and beyond spontaneous improvement. The decades following the initial empirical debate was characterized by a steady flow of evidence, aggregated with meta-analyses, supporting the claim that psychotherapy is indeed efficacious when compared to a wait-list condition (Lipsey & Wilson, 1993; Shapiro & Shapiro, 1982; Smith & Glass, 1977). The initial wave of meta-analyses has later been critiqued for not accounting for a range of problems that can inflate treatment effects. Specifically, *publication bias* (Staines & Cleland, 2007) entails that researchers shelf poor treatment results and publish successes thereby skewing results. Also, aggregating different treatments for different disorders can entail an *apples and oranges* problem (Philips, 2009) where aggregate summaries are made between different processes that are distinct and heterogeneous in nature. Another potential issue not managed by early meta-analyses was *researcher allegiance* (Munder et al., 2013). This is when a researcher is invested in the outcome of a study or has prior beliefs regarding the superiority of the treatment that might bias data interpretation.

More recent meta-analyses have attempted to compensate for these methodological challenges with statistical tools that allow for more flexible modeling of treatment data. These modern meta-analytic syntheses demonstrate smaller overall effect sizes, but nevertheless conclude that psychotherapy is efficacious (Cuijpers et al., 2010; Newby et al., 2015). For instance, when applying statistical control for publication bias and only selecting low-bias studies, Cuijpers et al. (2010) found an overall effect size of Cohen's $d = .42$ for the psychotherapeutic treatment of depression compared with various control group conditions. This effect size is illustrated below in figure 1.

Figure 1: Overlapping Distributions Cohen's *d* = .42



1.2 The Great Psychotherapy Debate

In many ways, the initial debate between Eysenck, Rosenzweig, and Strupp set the stage for what Wampold (2015) later has dubbed "the great psychotherapy debate". The controversies constituting this debate are numerous and cover issues of methodology, theory, epistemology, and the nature of psychotherapy itself. Are some therapies more effective than others? Does therapy require specific interventions or procedures in order to be efficacious? Should outcomes studies focus on quantifying improvements with standardized questionnaires or are improvements idiosyncratic in nature which requires qualitative assessments? Should psychological treatments be based on a scientific understanding of psychology and human nature and how should we define science (Rieken & Gelo, 2015)?

The two sides of this debate roughly correspond to the division of a *common factor* and *specific effect* view of psychotherapy. Common factors are general healing processes that are present across different treatment modalities, according to Frank (1973). Common factors include the expectation of getting help, a therapeutic relationship, a rationale for treatment, a ritual for its alleviation, and active involvement from both therapist and patient. The

proponents of the common factor model consider psychotherapy a culturally situated healing practice. They reject the notion that the process can, or should, be reduced to discrete interventions for specific pathological processes as proposed by proponents of the *specific effects* view of psychotherapy. Examples of proposed specific effects include fear extinction through systematic exposure of phobic stimuli (Abramowitz et al., 2019), planned behavioral activation to combat inactivity in depression (Kanter & Puspitasari, 2016), sleep deprivation therapy for insomnia (Taylor & Pruiksma, 2014) and the empty-chair technique to transform maladaptive emotions into adaptive ones (Elliott et al., 2004).

Another controversy linked to the great psychotherapy debate is the role of empirically supported treatments. The term originated with an APA task force charged with assessing which treatments had empirical support. One explicit goal of the task force was to secure and bolster psychotherapy's place as a valid treatment option in competition with interventions delivered by a psychiatrist (Chambless, 1995). The task force conclusion rekindled the debates of old and continue to influence how therapy is measured and practiced (Tolin et al., 2015). Perhaps the most prominent example of this influence is the development and implementation of the NICE (Clark, 2011) guidelines in the U.K. and the turn towards evidence-based treatments in the U.S. Veterans Health Administration (Karlin & Cross, 2014).

Although the common factor and specific effects view can be juxtaposed, they can also be construed as compatible perspectives on a multilayered psychotherapeutic process. Many have argued that a forced dichotomization has served the field poorly and new syntheses have been proposed (Weinberger, 2014). Others have argued for treatment guidelines and training programs that not only reflects traditional specific interventions but also include common factor elements such as the establishment of a working alliance, facilitating an environment of trust, and motivating patients for change (J. Brown, 2015).

The relative importance of common factor training rests on a series of observations: First, evidence indicates that therapists do not improve with experience (Tracey et al., 2014) even though many professional psychotherapists continue to receive technique-specific intervention training throughout their career (Bernhard & Goodyear, 2019). Second, there is a vast research literature indicating that the quality of the working alliance between therapist and patient is conducive to better treatment outcomes (Flückiger et al., 2018). Third, there is scant empirical documentation that indicates that learning specific interventions is essential for good outcomes (Bernhard & Goodyear, 2019). Finally, so-called "placebo treatment", which arguably only delivers common factor interventions have also been shown to be efficacious for many disorders (Kirsch et al., 2016). As these discussions highlight, the debate over the relative importance of specific techniques and common factors lives on. In a sobering analysis, Cuijpers et al. (2019) highlight the profound methodological challenge separating the causal dynamics between common factors and specific technique influences and conclude that no such research exists today.

1.3 Distinguishing Psychotherapy Efficacy and Effectiveness

In a special section of the Journal of Consulting and Clinical Psychology, David Barlow (1981) wrote that "at present, clinical research has little or no influence on clinical practice. This state of affairs should be particularly distressing to a discipline whose goal over the last 30 years has been to produce professionals who would integrate the methods of science with clinical practice to produce new knowledge" (p. 147). Following Barlow's statement, many researchers oriented themselves toward the clinical utility of research. For outcome studies, this shift entailed a focus on *effectiveness* trials with a design that included the specific naturalistic conditions present in standard clinical practice. This is in contrast to *efficacy* trials that are targeted to answer whether a particular treatment works when assessed under carefully controlled conditions, where treatment and patients can be monitored and

continuously assessed. The main issue was no longer whether psychotherapy worked in the lab, but whether it works in the traditional health-care system (Wallerstein, 2001). The topic of psychotherapy effectiveness was also called into attention by the conclusions of Weisz et al. (1992) meta-analysis of psychotherapy for children and adolescents. The report concluded that while there were several investigations of treatment outcomes from traditional clinics, these *did not* demonstrate statistically significant positive effects. These results stood in contrast to laboratory settings, where the positive treatment effects were evident. This initial result gave rise to increased awareness regarding the possible divergence between results obtained in the lab and in the field.

A set of later meta-analyses on psychotherapy interventions for adults concluded that outcomes under routine care operations closely resembled that of efficacy assessments (Shadish et al., 1997, 2000). These meta-analyses also explored the features that were likely to influence observed therapy effectiveness. One finding was that successful effectiveness trials often utilized disease-specific measures that were also targeted by the intervention itself, e.g. measuring the tendency to panic while treating panic disorder. In contrast, trials measuring global symptoms showed smaller effects. Another main finding was that successful trails often provided a higher therapy dosage compared to unsuccessful ones (Shadish et al., 2000).

The main finding from the psychotherapy effectiveness literature is that therapy does indeed seem to work in a traditional clinic setting (McAleavey et al., 2019; Peeters et al., 2013). This conclusion rests on studies that benchmark RCT-results from a particular treatment modality, often cognitive behavioral therapy, with the results obtained from traditional clinical care. A benchmarking trial uses an aggregate measure of changes with which the results from clinical care are compared. This comparison can be made using a

heuristic approximate approach, where the researcher compares overall effect sizes, or formally using a statistical tool such as equivalence testing (Lakens, 2017).

The benefit of comparing treatments heuristically is the ease of implementation. One example of this approach is the comparison made by Wade et al. (1998) who compared effect sizes obtained from a community hospital center treating panic disorder with the results found in two RCTs. The authors concluded that overall effect sizes were within a reasonable threshold of equivalence and concluded that the treatment worked equally well across the two settings. The downside of the simple heuristic approach is that results can be difficult to interpret as they lack a formal statistical framework for evaluation. A formal equivalence test has the benefit of providing this framework. It allows for conclusions beyond the null-hypothesis paradigm (equivalent, not equivalent), as equivalence tests can conclude both larger, smaller, and no-difference in effect. This approach can be seen in Minami et al.'s (2008) investigation of psychotherapy for depression in a managed care environment. This investigation compared the results from several investigations of psychotherapy for depression RCTs against a record of 5703 patients treated in local treatment facilities. They concluded that results were equivalent between the two settings.

Although overall results from benchmarking trials are positive regarding the effectiveness of psychotherapy in a routine setting, some prominent caveats should be noted. First, the benchmarking literature is sparse and does not cover the entire spectrum of mental illnesses found in a typical psychiatric clinic. There are no benchmarking investigations of psychotherapy for personality disorders or including patients with psychotic disorders for instance. Also, what we appraise as a traditional psychotherapy clinic or treatment-as-usual is a diverse and heterogenic set of different treatment conditions. It is reasonable to assume that there are important differences in treatment outcomes across different clinics, serving different patient cohorts, under different operating parameters. Today, the majority of

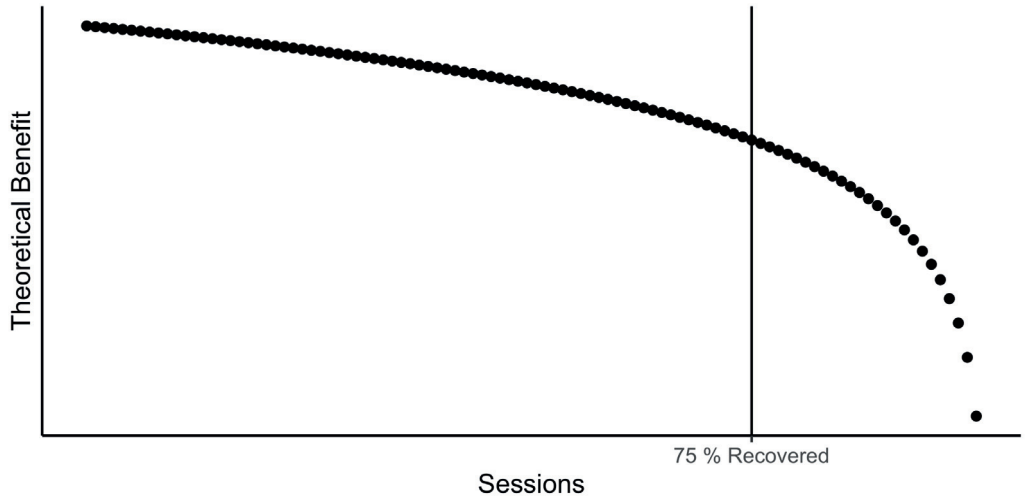
effectiveness data utilized in the benchmarking literature comes from psychotherapy conducted at university counseling centers (McAleavey et al., 2019). The problems with exclusively investigating student samples have been noted by researchers in many fields of psychology, including clinical research (Henry, 2008).

1.4 Psychotherapy Dosage

There is an extensive research literature dedicated to exploring how psychotherapy dosage relates to outcomes. The answer to the question of dosage can have a considerable impact on psychotherapy practice as it relates to the normative evaluation of how much therapy should be supplied by a private insurer or a governmental programmed tasked with the responsibility of providing healthcare. If a particular treatment length provides a maximum expectable benefit, then this observation can be used as an argument for the normative decision that patients should be allocated treatment of this length. The view of psychotherapy dosage as a single optimal value parallels pharmaceutical dosages where medicine is usually prescribed in a particular dosage for a particular disorder (Kopta, 2003).

The focus on psychotherapy dosage was kickstarted with the work of Howard et al. (1986) who provided the first formal analyses of psychotherapy efficacy in relation to the length of therapy. The authors aggregated data from previous investigations and coded treatment outcomes in terms of a dichotomous recovered/not recovered thresholds. The patient sample were gathered from American university counseling centers (n = 827), private psychiatric clinics (n = 593) and community outpatient clinics (n = 1159). These results were combined and assessed using probit regression which demonstrated a negatively accelerating curve (log-linear function) of proportion recovered per session in therapy. Such a log-linear function is illustrated in figure 2.

Figure 2: Theoretical Model of Logarithmic Psychotherapy Benefits



This indicated that the majority (75 %) of patients recovered after 26 sessions. This early work also stressed that patients in different diagnostic categories demonstrated different patterns of recovery (Howard et al., 1986). E.g., patients diagnosed with borderline-psychotic conditions required significantly more therapy to achieve recovery compared with patients with depression or anxiety.

This early investigation energized the ongoing debate of psychotherapy dosage (Kolden, 1991). This controversy was arguably fueled by the recommendation of Howard et al. (1986) that a fixed dosage of 26 sessions as a "rational time limit" (p. 163) which could be considered a rule-of-thumb for maximum dosage. Following this lead were several investigations of psychotherapy dosage at different treatment sites. Most investigations used data from university counseling clinics and find "optimal doses" ranging from 4 -11 sessions (Anderson & Lambert, 2001; Baldwin et al., 2009; Draper et al., 2002). However, several investigations have also shown that more severely afflicted patient populations require far larger doses in order to achieve recovered status. Asay et al. (2002) found that a group of patients with characterological pathology (N = 29) required between 42 – 54 sessions in order to reach the hypothetical point of diminishing returns. Similarly, Lincoln et al. (2016) assessed outcomes from an outpatient clinic treating psychosis (N = 58), concluding that the

group demonstrated diminishing returns after 25-30 sessions. In a recently published review of dose-effect investigations of naturalistic psychotherapy, the authors concluded that 95 % of patients that have been assessed are treated at university counseling centers (Robinson et al., 2019).

1.5 Long-Term Psychotherapy Research

The brevity of the psychotherapy dosage literature stands in contrast to investigations on long-term psychotherapy. There is no clear-cut definition of long-term psychotherapy, but it is commonly associated with either psychodynamic or psychoanalytic treatment with few exemptions. The focus of long-term psychotherapy treatment is often relational processes that are explored using transference, support, and interpretive interventions (Gabbard, 2017). There is also no formal definition regarding the minimum length of long-term psychotherapy but meta-analyses commonly implement a minimum of 40 sessions or one year of treatment (Abbass et al., 2014; Leichsenring & Rabung, 2011). Also, treatments are usually open-ended in contrast to fixed treatment lengths. There is a substantial empirical literature supporting the efficacy of long-term psychodynamic or psychoanalytic treatment (Leichsenring & Rabung, 2011; Shedler, 2010; Steinert, Munder, et al., 2017; Woll & Schönbrodt, 2020). These findings indicate that long-term treatments are associated with equal effect sizes compared with evidence-based treatments. These findings are in line with the overall discovery that different psychotherapy therapies tend to produce similar outcomes, known as the dodo bird verdict (Budd & Hughes, 2009).

Some proponents of long-term psychotherapy argue that this approach is superior when treating patients with complex psychopathology (Woll & Schönbrodt, 2020). This assertion is often based on dose-response research that indicates that patients with more severe disorders require longer treatments. Similarly, it has been hypothesized that a longer treatment will result in higher retention of gains, compared to short treatments (Perry & Bond,

2009). Another feature of research on long-term psychotherapy is the utilization of a cohort study design, in contrast to an RCT. A cohort study assesses a group of treated patients but does not control the treatment effect with a control condition. Maat et al. (2007) argue that the RCT format is unsuitable for assessing psychotherapy effects because the control condition is not feasible when patients are severely afflicted with a mental illness. The authors argue that no-treatment, wait-list, or placebo control can be both unethical and practically impossible.

Similarly, it is argued that the treatment-as-usual condition, heavily utilized by research on the treatment of personality disorder, is seldom comprised of anything other than no-treatment or psychopharmacology. Maat et al. (2007) conclude that "RCTs represent not the highest but rather an irrelevant level of empirical evidence. Where RCTs are an impracticable and, therefore, inadequate research method, cohort studies provide the best available evidence." (p. 64). Several researchers and clinicians have raised their voices against an RCT-based movement towards evidence-based treatment. Critics of this movement highlight the value of cohort and case studies for their high ecological validity and contextual nuance (Milton, 2002). Defenders of the traditional RCT-framework, on the other hand, point out that randomized conditions with experimental control are the only way to satisfy demands of interval validity, which is argued to hold primacy over external validity (D. T. Campbell, 1986). They also accuse the critics of the RCT-paradigm of committing the nirvana fallacy, that is, rejecting a beneficial approach because it does not perfectly attain its goals (Lilienfeld et al., 2018). For example, few would argue that the Beck Depression Inventory perfectly captures every dimension of interest in a depression trial, but equally few would argue that a mean score tells us nothing of interest.

Also, while wait-list control has been shown to be inexact control condition (Patterson et al., 2016), few would argue that this imperfection should lead to the discarding of all research where it is implemented. Lilienfeld et al. (2018) summarize their defense of the

RCT-methodology with an adapted quote from Winston Churchill quip on democracy: "RCT's are the worst outcome designs except for all other designs that have been tried." (p. 537). In many ways, this ongoing debate is the modern version of *the great psychotherapy debate* (Wampold, 2015) initiated by Eysenck, Rosenthal, and Strupp.

The assertion that long-term psychotherapy produces superior outcomes for complex cases has been investigated by a select few RCTs which randomized patients to either long-term psychotherapy or an evidence-based treatment such as CBT. The Helsinki Psychotherapy Study randomized 326 out-patients with mood and anxiety disorders to either long-term psychodynamic, short-term psychodynamic, or solution-focused therapy. Patients were recruited from a mix of sources including routine outpatient psychiatric clinics, private practitioners, as well as student and community clinics. To be included, patients had to have experienced long-standing (>1 year) depressive or anxiety disorders that caused work dysfunction. Patients with a psychotic disorder, severe personality disorder, adjustment disorder, bipolar disorder, or substance abuse were excluded. At pretreatment, 18.1 % qualified for a personality disorder while 42.9 % had psychiatric co-morbidity.

The researcher assessed outcomes for several years following treatment completion with the last assessment 10 years following treatment. The results demonstrated that long-term psychotherapy produced statistically significant superior outcomes compared to short-term psychodynamic and solution-focused therapy (Knekt et al., 2016). However, this effect was modest while requiring far more treatment. The median treatment length was 192, 142, and 60 sessions for long-term psychodynamic, short-term psychodynamic, and solution-focused therapy, respectively. The authors conclude that short-term treatments could arguably establish similar treatment effects by adding an option of auxiliary treatment for patients in need. They also found that a greater proportion of patients in the shorter treatments pursued

auxiliary psychotherapy compared to long-term treatment (short-term psychodynamic: 58 %, solution-focused: 47 %, long-term psychodynamic: 33 %).

Similarly, Leuzinger-Bohleber et al. (2019) randomized 252 adults with chronic depression into either long-term CBT or long-term psychoanalytic therapy. The mean treatment length for the CBT group was 57, in comparison to psychoanalytic therapy which was 234. Patients had to be depressed for more than 1 year and meet diagnostic criteria of major depressive episode or dysthymia. The majority of patients had one (25 %) or two or more (45 %) previous unsuccessful treatment attempts. Combined, the treatment groups demonstrated large symptomatic improvements (Cohen's $d = 1.83$) at a three-year follow-up. Crucially, no statistically significant differences were found between the CBT and psychoanalytic therapy conditions.

1.6 Personality Disorders in Outpatient Treatment

Having a comorbid, or primary personality disorder is commonplace in a psychiatric outpatient clinic setting. Pinpointing an exact epidemiological estimate is challenging because of the variability in the disorder, the use of unstandardized diagnostic tools, and the challenge of demarcating when problems are fittingly severe to qualify for a personality disorder. Therefore, epidemiological study estimates ranging from 30 – 80 % (Beckwith et al., 2014). The American Diagnostic and Statistical Manual of mental disorders (4th ed.) categorize personality disorders into three clusters (American Psychiatric Association, 1994). Cluster A personality disorders are characterized by odd and eccentric behavior that others can find peculiar, odd, or suspicious. It includes Schizotypal, Paranoid, and Schizoid personality disorder. Cluster B personality disorders are characterized by dramatic and impulsive behaviors while experiencing very intense emotional instability. It includes borderline, histrionic, antisocial, and narcissistic personality disorder. Lastly, Cluster C personality

disorder is characterized by pervasive anxiety and fearfulness. It includes dependent, obsessive-compulsive, and avoidant personality disorder.

There are few epidemiological studies of personality disorders compared with studies of clinical disorders. Torgersen et al. (2001) assessed a representative community sample of N = 2053 utilizing the structured interview for the DSM-III-R and found prevalence rates of 4.1 % (Cluster A), 3.1 % (Cluster B) and 9.4 % (Cluster C). There is also sparse empirical literature on the treatment of personality disorders, compared with other mental disorders. In their review, Bateman et al. (2015) conclude that the majority of studies have small sample sizes, little control of comorbidity, and a short or non-existing follow-up assessment. Also, the research is focused almost exclusively on the treatment of borderline personality disorder with very little documentation on other personality disorders. However, the available evidence on borderline personality disorder is largely optimistic about psychotherapies' ability to alleviate or rehabilitate personality disorder.

The majority of this research is on specialized interventions that are tailored for the treatment of borderline personality disorder, such as dialectic behavioral therapy, mentalization-based therapy, transference focused therapy, schema therapy, or cognitive analytic therapy. A recent review found that specialized therapies demonstrate superior outcomes compared to treatment-as-usual or wait-list control with no discernable difference between treatment types (Oud et al., 2018). The same review also documents that there is great heterogeneity in treatment effects, a high degree of drop-out, and a high risk of researcher allegiance bias.

2. Aims of the Thesis

The overall objective of this thesis is to explore outcomes from an investigation of patients treated with open-ended psychotherapy under naturalistic, clinically representative conditions. This entails supplementing the psychotherapy treatment literature with an empirical investigation that closely resembles the traditional outpatient clinic with a highly heterogenic patient sample, including a substantial proportion with severe mental disorders. This is in contrast to the traditional RCT format which seeks to employ experimental control in order to increase the validity of specific causal claims. The internal validity demonstrated by the typical RCT psychotherapy investigation is, however, at risk of reducing external validity as results might differ in a naturalistic setting with less stringent treatment delivery. The treatment data explored in this thesis demonstrates some key features that increase its external validity. Firstly, for the majority of therapists, no restrictions were placed on what type of psychotherapy methodology or interventions they could choose. Instead, therapists were free to implement their practice as they best saw fit. Secondly, the majority of treatments were open-ended and could be specifically tailored to the patients' needs in terms of the frequency of sessions. Patients and therapists were instructed to come to a mutual agreement regarding the length and intensity of treatment. Thirdly, the patient sample was highly heterogeneous including mild, moderate, and severe cases of psychopathology. A specific goal of this thesis is to explore psychotherapy outcomes including patients with moderate to severe characterological psychopathology. For this reason, patients were recruited so that approximately half of the sample was diagnosed with a personality disorder at pretreatment. Another goal of the thesis is to analyze outcomes during psychotherapy, but also include a two-and-a-half-year follow-up period in order to assess whether changes are stable in the long run after treatment ends. This combination of a diverse patient cohort, high variability in

treatment lengths, rigorous measures of psychopathology, and a considerable follow-up period is rare in the psychotherapy outcome literature.

2.1 Paper I

The goal of paper I is to explore our results from the main outcomes measures including both self-report and observer-rated measures. These analyses document the extent of change during treatment, the proportion of patients that achieve reliable and clinically significant change, as well as how many patients demonstrate reliable deterioration. Another aim of paper I is to examine the changes in observer-rated symptom and personality disorder diagnoses following treatment and in the follow-up period. Paper I also aim to compare and contrast self- and observer-evaluated assessments. The results of these analyses can highlight the possible convergence or divergence between how patients evaluate themselves and the corresponding diagnostic evaluation of a trained, clinical professional. The investigation of deterioration is particularly of interest as there are few investigations of this phenomenon under open-ended and personalized treatment conditions. It has been proposed that flexibility in treatment conditions can influence levels of patient deterioration (Rozental et al., 2018). Analyses of deterioration rates in this study can help clarify the empirical basis of this proposition.

Another goal is to assess changes in occupational functioning following treatment and in the two-and-a-half-year follow-up period. The research of return-to-work and mental health suggests that interventions that solely focus on the treatment of mental illness rarely produces improvements in occupational functioning (Cullen et al., 2018). However, this research is heterogeneous in terms of the interventions supplied with the majority consisting of manualized, short-term CBT. The results of paper I can help clarify the relationship between mental health interventions and occupational functioning when the treatment is personalized in terms of intensity and duration.

2.2 Paper II

The goal of paper II is to subdivide overall outcomes into two main treatment conditions, namely patients with and without a personality disorder. The paper aims to answer whether patterns and magnitudes of change over time in psychotherapy differ for patients with and without a PD. This is arguably a salient diagnostic distinction with a relatively meager research literature. Previous research is unequivocal regarding the proposition that patients with a personality disorder demonstrate poorer outcomes compared to patients without when undergoing comparable treatments. The few investigations that exist mainly consist of manualized CBT interventions. It seems plausible that treatment flexibility, in terms of intensity and duration, and treatment lengths could play a moderating role in the relationship between outcomes and personality disorder. In addition to a dichotomous conceptualization of personality disorder, paper II also aims to investigate this relationship utilizing a continuous measure of the magnitude of personality problems. Assessing the magnitude of personality problems with a continuous measure holds the promise of a more nuanced approach that can go beyond the validity limitations of a dichotomous conceptualization (Furnham et al., 2014). Paper II aims to explore whether the magnitude of personality disorder problems compared with a dichotomous conceptualization influences how it affects treatment outcomes. Paper II thus also investigates whether patients with and without a personality disorder show different rates of retention of gains.

2.3 Paper III

The aim of paper III is to assess the relationship between psychopathology severity and the rate and magnitude of change. This relationship will be explored utilizing a heterogeneous patient cohort that has received highly varying psychotherapy treatment dosages. The variability in treatment dosage and levels of psychopathology has the potential to explore a far greater range of treatment lengths compared to the majority of the

psychotherapy dosage literature which rarely exceeds 20 sessions. The paper will address the question of psychotherapy dosage from an open-ended treatment point-of-view, which is rare in the psychotherapy dosage literature. Another aim is to assess the overall pattern of change and investigate what statistical model provides the best fit of the observed outcomes. Early work in this field suggested that patient's improvements follow a log-linear rate. That is, the benefits of psychotherapy are substantial at the beginning of therapy with a following reduction in effectiveness as treatment progresses. This is in contrast to a linear model where the rate of change is sustained throughout treatment. We expect our results to differ from what is found in the majority of the psychotherapy literature, as it is mostly comprised of short treatments and patients with less severe psychopathology.

3. Methods

3.1 The Norwegian Multicenter Study of Process and Outcomes in Psychotherapy

All treatment data presented in this dissertation originates with the Norwegian Multicenter Study of Process and Outcomes in Psychotherapy (NMSPOP). The NMSPOP is a naturalistic treatment study with patients (N=370) gathered from eight treatment sites within the Norwegian public health system in the years 1995 – 2008. The majority of patients (n = 301) were recruited from 17 separate psychiatric outpatient clinics. Each clinic belongs to one of the following six Norwegian regional hospitals: Telemark Hospital Trust, Oslo University Hospital, St. Olav Hospital Trust, Southern and Eastern Norway Regional Hospital, Innlandet Hospital Trust, and Vestfold Hospital Trust. A subsample of data was gathered data from the Norwegian University of Science and Technology's student training clinic (patient n = 27) located in Trondheim. Lastly, we gathered data from several outpatient clinics with physiotherapists (patient n = 42) undergoing specialization in psychodynamic body therapy for patients with somatoform disorders. Therapists and patients from this source were dispersed across a range of clinics across Norway.

At each of the eight sites, trained coordinators (clinical psychologists or psychiatrists) were responsible for recruiting patients and administering the research protocol. We instructed the coordinators to select patients from their local population at random while ensuring that approximately half had a diagnosable personality disorder. We did not apply any formal randomization procedure. The local coordinators also assessed the patients. The coordinators were all experienced clinicians who underwent training in the use of the assessment instruments. The inclusion policy was liberal, with the following exclusion criteria: age less than 20 years, active psychosis, drug/alcohol abuse as the primary problem, need for emergency treatment or hospitalization, and mental retardation (IQ < 70). These criteria are in line with commonly used criteria in the evaluation of patients for individual

psychotherapy at outpatient clinics. The Regional Committee for Medical Research Ethics in Eastern Norway approved the protocol for the study.

After receiving information and signing a written consent, the patients were submitted to a two-step pretreatment assessment. In the first step, patients completed several self-report questionnaires, including among others a sociodemographic inventory, questions regarding occupational functioning, psychiatric symptoms, and levels of interpersonal problems. In the second step, patients underwent a structured diagnostic assessment by the coordinator at each site. This assessment comprised of a Structured Clinical Interview (SCID) based on the Diagnostic and Statistical Manual of Mental Disorders 4th edition (American Psychiatric Association, 1994) criteria for Axis I and II disorders. The patients were assigned to therapists based on availability after the initial assessment. Patients completed self-report questionnaires during treatment after the 3rd, 12th, and 20th session. Patients then completed self-report questionnaires every 20th session following the 20th session for as long as they received therapy. Following treatment completion, the coordinator repeated the diagnostic evaluation with a SCID I and II interview. The self-report questionnaires were also repeated at the posttreatment assessment. While some of the patients completed their post-assessments directly after treatment completion (35 %), most completed this assessment a few months after treatment completion due to practical issues. The median delay was 2.5 months after treatment completion (SD = 25.5, Mean = 9.8). The patients were then assessed with SCID interviews by a coordinator and completed self-report measures six months, one, and two and a half years following the posttreatment assessment. A subsample (n = 17) of patients also had a six-year follow-up assessment.

3.2 Measures

The Symptom Checklist 90 Revised (SLC-90-R). We used the SLC-90-R (Derogatis & Unger, 2010) to assess overall symptom presence and severity. It contains 90 questions

asking patients to rate, on a Likert scale from 0 (not at all) to 4 (very much), the intensity of a given symptom during the last week. Items span a wide range of mental health problems including subjective bodily complaints, problems related to interpersonal functioning, psychotic or hallucinogenic symptoms, symptoms of anxiety and depression, as well as cognitive deficits such as problems with executive functioning, attention, and memory. We used the Global Severity Index (GSI) as our overall symptom severity measure, which is the mean rating across the entire checklist.

The Inventory of Interpersonal Problems 64 (IIP-64). We applied the IIP- 64 (Horowitz et al., 1979) to assess levels of interpersonal problems. It consists of 64 questions rated on a five-point Likert scale from 0 (*not at all*) to 4 (*very much*). The first 39 questions begin with the phrase "*It is hard for me to…*" while the remaining 25 questions ask about "*Things that I do too much*". The interpersonal problems that are measured span the range of DSM IV-R personality disorder. We used the IIP Global score to obtain a single measure of interpersonal functioning, which is the mean across the entire inventory.

Structured Clinical Interview for DSM-IV Axis I & II Disorders (SCID I & SCID-II). The SCID interview was developed for the assessment of psychiatric disorders as defined by the DSM IV-R. The SCID I appraises all mental disorders with the exception of intellectual disability and personality disorders while the SCID II exclusively appraises personality disorders. Each condition is associated with a set of criteria that constitute different aspects or manifestations of the disorder. The criteria are assessed through open-ended questions or items across the various diagnoses. Each question or item can be scored as either absent, sub-threshold, true, or "inadequate information to code." The criterion items corresponding to each particular condition is then summed to assess whether the patient qualifies for a given disorder. The number of positive SCID II items can also be used as a continuous measure of the severity of characterological or personality problems.

Occupational status. We created a dichotomous variable (occupational functioning vs. no functioning) based on self-reported occupational status. We classified the following responses as "functioning": 1) "I am currently engaged in paid work," 2) "I am a stay-at-home mom/dad," 3) "I am currently engaged as a student" or 4) "I am retired." We classified "non-functioning" with the following responses: 1) "I am currently on sick leave," 2) "I am currently in work rehabilitation," 3) "I am currently receiving disability benefit" or 4) "I am currently unemployed." We assessed changes in occupational status by comparing pre- to posttreatment levels and pre- to follow-up measurements.

3.3 Patient and Therapist Characteristics

Overall, patients were highly heterogeneous in terms of demographics and severity of mental illness. The mean sample age was 35.2 (SD = 9.4) years with a majority of female patients (69.5 %). The mean number of pretreatment SCID II criteria of personality disorder was 12.7 (SD = 5.8). The pretreatment mean symptom score, as measured by the Global Severity Index (GSI) of the SCL-90-R, was 1.28 (SD = 0.61). This is a similar level of distress to that reported by other investigations of outpatient clinics (e.g. Langeland et al., 2006). We found a mean rating of interpersonal problems (IIP Global), as assessed by the IIP-64, of 1.49 (SD = 0.52). This level of interpersonal problems also reflects a typical level of interpersonal problems in an outpatient setting (Bjerke et al., 2011).

Axis I disorders were distributed as follows: affective, 50%; anxiety, 64%; somatoform, 24%; eating-disorder, 9%; substance-related, 3%; other, 5%. In total, 86.8 % of patients had one or more Axis I disorders. Distribution of Axis II disorders were as follows: Cluster A, 22%; Cluster B, 19%; Cluster C, 46%; Not Otherwise Specified, 2%. In total, 54 % of patients had at least one personality disorder. The mean number of pretreatment SCID II criteria of personality disorder was 12.7 (SD = 5.8). The patients reported that "the problem which you are now seeking treatment for" had lasted on average 11.7 years (SD = 9.75). The

sample did not include patients who sought treatment for a primary substance use diagnosis.

When assessed pretreatment, some patients indicated that they used prescribed medication for their psychological problems either “regularly” (22 %) or “when in need” (7 %). The majority of patients using psychotropic medication indicated that they mainly used an antidepressant (n = 72) while fewer indicated that they mainly used an anxiolytic (n = 19), a hypnotic (n = 3), an antipsychotic (n = 4), or pain medication (n = 8). Of the psychotropic medication users, the majority used a single medication (n = 70), while some were prescribed two (n = 22), three (n = 9) or four (n = 4) concurrent medications.

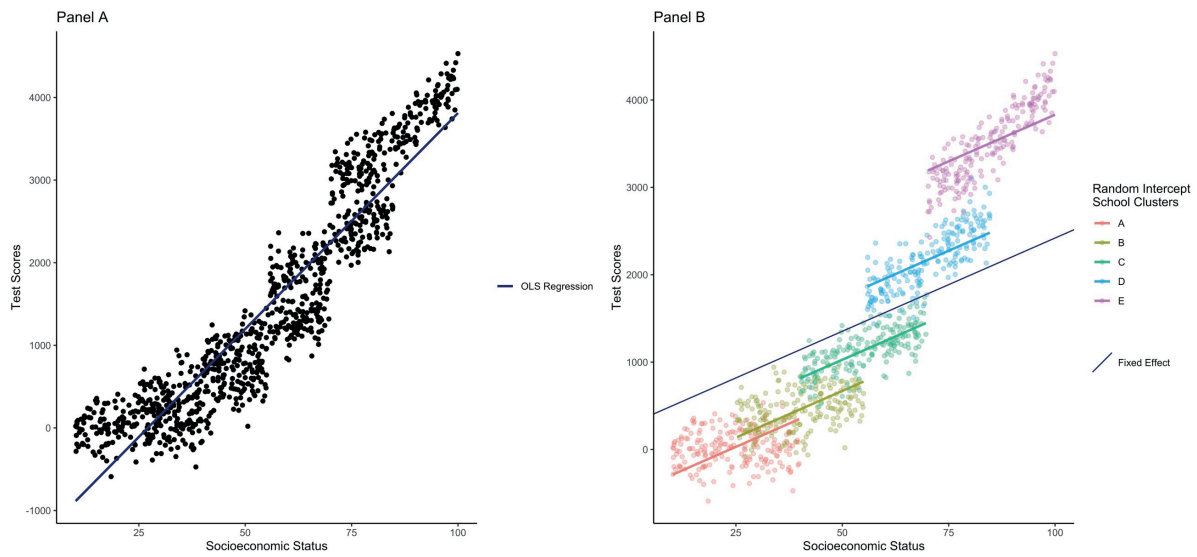
The therapists (n = 88) were mainly experienced clinicians with a mean of 10 years (SD = 6.5) of psychotherapy experience. The therapist also had postgraduate professional training, including a mean of 5.9 years (SD = 4.3) of clinical supervision. Notable exceptions were the physiotherapists (n = 8) and student therapists (n = 27) who received close supervision. The mean number of patients per therapist was 5.6 excluding the student therapists where each student saw one patient. Therapists were instructed to provide their usual therapeutic practice.

3.4 Statistical Analyses

The main research questions presented in this thesis rests on analyses performed using the Multi-Level Model (MLM) framework. The MLM framework is an extension of traditional ordinary least-squares (OLS) multiple regression that allows for the specification of both *random* and *fixed* coefficients, corresponding to different levels of analysis (Snijders & Bosker, 2012). In practice, it can be difficult to separate what effects correspond to random vs. fixed (Gelman, 2005). A simple heuristic is that an effect is fixed when it is of interest in itself, while random effects are only of interest to understand an underlying population (Searle et al., 1992). The MLM framework was developed as an analytic tool for data that contained multiple sources of variability which corresponded to a hierarchical data structure. One of the

earliest applications of MLM was on educational research testing hypotheses across different schools (Burstein, 1980). The hierarchical nature of this data implies dependencies in the data structure, as children from the same school or classroom would be more similar compared to children from different schools or classrooms. These dependencies break with the independence assumption in traditional multiple regression. A traditional OLS analysis using data with dependencies can result in the erroneous estimation of both standard errors and coefficients (Bryk & Raudenbush, 1987). Figure 3 illustrates these dependencies as different schools together in a simulated dataset measuring a test score and Socioeconomic Status (SES).

Figure 3: OLS and Multi-Level Model Comparison



Panel A represents a standard OLS regression assessment of test scores and SES using the formula shown below:

$$Y = \beta 0 + \beta 1 + \ e$$

This expression assumes that the observations that are sampled are independent of each other. This assumption is violated in a school setting as children in similar schools will likely have similarities that can be ascribed to causal influences that follow the school clusters (e.g. socioeconomic status, school environment, teacher competencies). Panel B shows the same

data but also demonstrates the pattern of clustering or dependency. MLM supplements the estimation of the overall linear effect with parameters that capture within-cluster effects. Panel B demonstrates this using a random intercept parameter that captures some of the between-school variance. The subscript i corresponds to the level of the individual while j corresponds to the school level.

$$Y_{ij} = \beta 0 + \beta 1 x_{ij} + u_j + e_{ij}$$

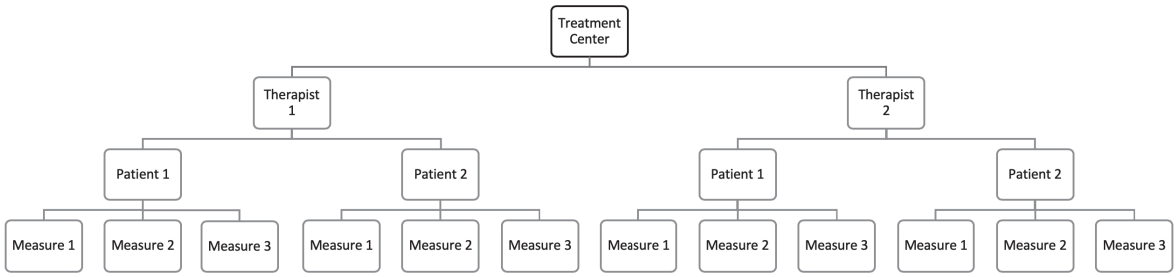
The variance in intercept between schools is captured by a single u_j variance parameter that captures the clustering effect and allows for a more precise estimation of the fixed effect not accounted for by clustering. The fixed and random part of the equation corresponds to the different levels of analysis. Notice that the fixed effect slope of school SES on test scores decreases as the model incorporates the variance explained by level 2 clustering.

Level 1: Fixed Effects $Y_i = \beta 0 + \beta 1 x_i + u_j + e_i$

Level 2: Random Effects $Y_{ij} = \beta 0 + u_j + e_{ij}$

The hallmark of a longitudinal dataset is that the same individuals are assessed across time and this represents a clustering-problem similar to educational data. The unit of dependency is mainly the individual. The assumption of independent sampling is violated as one cannot assume that an individual will supply uncorrelated data across time. On the contrary, one would expect the exact opposite, namely that individuals show some stability across time (Steele, 2008). Figure 4 illustrates the nested data structure that is relevant for psychotherapy trials.

Figure 4: Nested Psychotherapy Data



Each individual measurement is nested within a patient, which is also nested within a therapist and the overall treatment center. The clustering effect within-individual is usually very high in longitudinal designs. Similarly, simulations studies indicate that neglecting the therapist effect can bias hypothesis testing (Magnusson et al., 2018). Previous work on the NMSPOP has already investigated the therapist effect (Nissen-Lie et al., 2013). The therapist effect was found to be 4.2 % and 2 % for the GSI and IIP-Global respectively. However, the multi-level models presented in paper I, II and III do not incorporate a separate therapist effect as it did not significantly improve the fit of the models. In addition to patient and therapist clustering, this dataset also has potential clustering at the treatment center level. The treatment center clustering was also not included in the models presented in paper I, II or III as the models would not converge with many overlapping fixed effects. An example of this would be when analyzing personality disorder both as a fixed effect, as measured by SCID II items, and as a random factor as some centers include more patients with personality disorders than others. In contrast, running a empty model without fixed covariates reveals that treatment centers accounted for 3.5% and 2.3% percent of the total variance in GSI and IIP-Global scores, respectively.

3.4.1 Estimation, Shrinkage and Pooling in Multilevel Models. Researchers apply longitudinal analyses of psychotherapy when they are interested in changes across time. As previously demonstrated, these changes can be modeled as both on the level of the individual and the overall group level. The estimated random parameters can be used patient-level

growth curves that can both incorporate starting levels of pathology and the patients' unique development over time. There are many ways of estimating these parameters but the most commonly used are the Restricted Estimation Maximum Likelihood (REML) and Maximum Likelihood (ML). The goal of these estimation procedures is to provide a set of parameters that provides the highest likelihood of observing the data at hand. The process is done iteratively while exploring a multidimensional parameters space where the topology is determined by the likelihood of the particular parameter producing the data as it is observed. MLM can thus be assessed using model fit criteria. This can be accomplished by comparisons to a saturated model with a perfect fit, where each parameter corresponds to an observation. Model fit, or model deviance, corresponds to the degree of badness of fit. A commonly used deviance statistic is -*2LL*:

$$D = -2 * \ln\left(\frac{likelihood\ for\ saturated\ model}{likelihood\ for\ alternativ\ model}\right)$$

This statistic can be expanded into fit statistics that take various critical aspects of the estimation procedure into account. The deviance statistic is usually foregone in the psychotherapy research literature as it does not take model parsimony into account. A more specified model will always result in a lower deviancy and might thus introduce statistical overfitting by the unwanted modeling of stochastic noise (Cheng et al., 2010). Other deviance procedures have been developed to balance goodness of fit with model parsimony. The Akaike Information Criteria (AIC) starts with *D* but includes an addition of total numbers of parameters * 2, thereby punishing overfitting. Similarly, the Bayesian Information Criteria (BIC) starts with *D* but includes an addition of the log of n number of participants, multiplied by the number of parameters used in the alternative model. It is thereby sensitive to the number of parameters in combination with the amount of data available for estimation. Note that *n* denotes the total amount of observations, not the sample size.

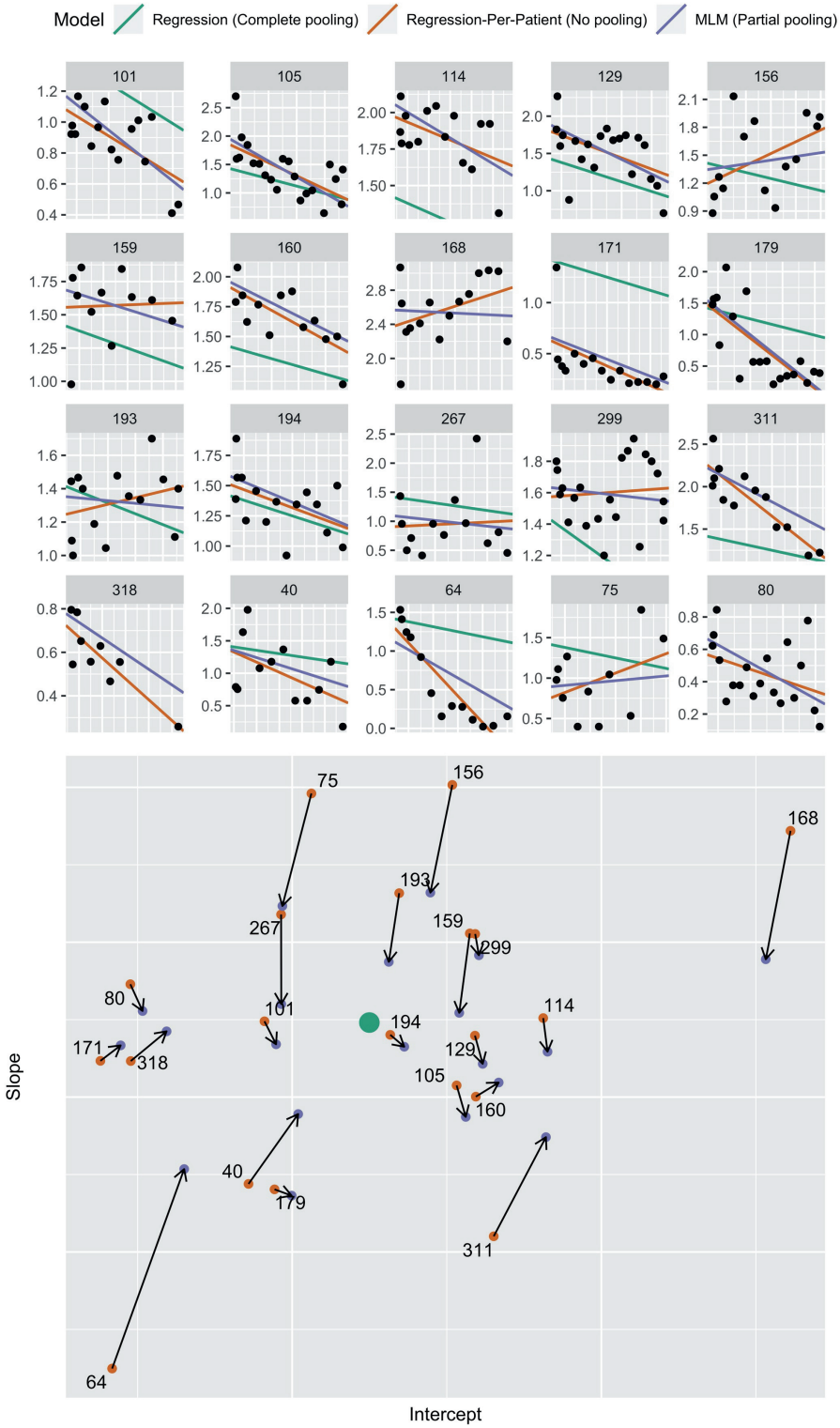
$$AIC = -2 * \ln\left(\frac{likelihood\ for\ saturated\ model}{likelihood\ for\ alternativ\ model}\right) + 2 * parameters$$

$$BIC = -2 * \ln\left(\frac{likelihood\ for\ saturated\ model}{likelihood\ for\ alternativ\ model}\right) + \ln(n) * parameters$$

The MLM framework provides tools for assessing both fixed and random parameters simultaneously. This is important as the two are interlinked sources of variation that can bias hypothesis testing and model construction. This disentangling is carried out by a regularization algorithm which strikes a tradeoff between variance and bias. In the case of figure 3 Panel A, the traditional OLS estimator provides the best fit by minimizing the sum of the squared errors, $\sum SSE$, but overfits the data due to encompassing school level variability. This reduces overall model generalizability.

In contrast, Panel B demonstrates regularization as the MLM `lmer` function (Bates et al., 2015) applies a penalized least squares regression which minimize the sum of the squared errors and adds a penalizing term, $\sum SSE + \lambda\ x\ the\ slope^2$. This procedure is calculated iteratively using cross validation of sections of the data analyzed separately. This shrinks the coefficient β as λ increases and variance between different subsections shrinks. After obtaining a penalized estimate of β, the algorithm then estimates random effect parameter by minimizing the sum of the negative log-likelihood. When applying REML estimation, data is initially transformed to cancel out fixed effect structure, thereby assuming its accuracy. This procedure produces less bias prone estimation of the random effect structure when there is sparse data. With large datasets, this effect is negligible (Gurka, 2006). However, if a researcher's main objective is to compare models with different fixed effects, then ML estimation is required.

Figure 5: MLM Parameters as Compared with OLS Regression



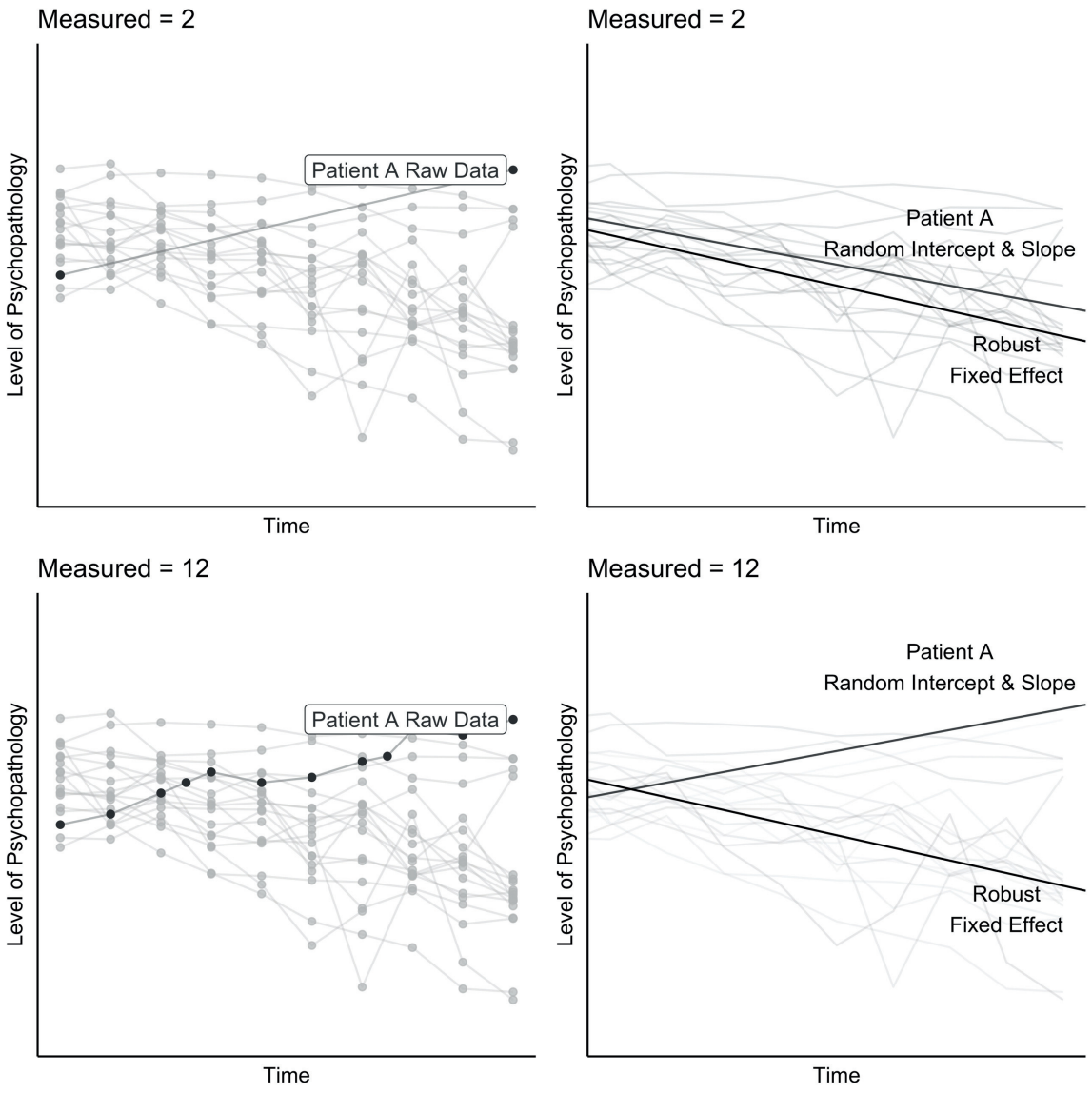
The features described above allow multi-level models to strike a balance between modeling each individual separately (No pooling) and modeling the entire sample (Complete pooling) without consideration for each individuals' idiosyncrasies. This is illustrated in figure

5. The top panel shows individual datapoints (black dots) for a sample of selected NMSPOP patients with long treatments. The x-axis represents time in treatment while y represents GSI scores. Each patient has been fitted with three models. Fitting unique regression parameters to each patient (Orange) does not pool any information from the rest of the sample, making these estimates highly variable. A single regression model (Green), fitted across the entire sample, represents complete pooling as all datapoints are summarized into a single slope and intercept parameter. In contrast, MLM (Purple) estimates allows for partial pooling that shrinks the individual estimates towards the overall fixed effect. The bottoms panel shows the same patients estimates with intercept (x-axis) and slope (y-axis) parameters. The arrows indicate how each estimate shrink the individual parameters towards the overall fixed parameter.

3.4.2 Using Random Effects to Model Patient Trajectories. There are currently two main approaches to modeling patient-level change trajectories utilized in the psychotherapy treatment literature (Laurenceau et al., 2007). One approach is to model change in terms of Latent Change Growth Models (LCGM) based on a structural equation modeling (SEM) framework. This entails creating latent growth factors that capture the patient level change variability. These factors are then utilized to model change as a function of an underlying growth process where the best-fitting parameters are equipped for each patient (Fitzmaurice et al., 2011). In contrast, MLM's assess development as a hierarchy of overall between patient variance and within-patient variance. Similar to LCGM, it does not measure the between patient variance directly but instead estimates this using the likelihood estimations procedures demonstrated above. This entails that a patient's predicted score is based on the variability in both levels of analysis, as well as the reliability of the patients' scores. Each patient is assumed to follow the overall group development, while diverging evidence can adjust the predicted patient scores. If the fixed effect demonstrates little variability across patients, then a unique patient-level effect needs several datapoints as evidence against the assumption that

the patient follows a similar trajectory as the overall group mean, or fixed effect. Figure 6 illustrates this by modeling change with both a fixed effect of time and random slope and intercept parameter.

Figure 6: Random Slope and Intercept Model with Varying Reliability.



The top panels illustrate how a patient with a diverging trajectory will get weighted towards the overall fixed effect. This is because the patient is only measured twice indicating low reliability. This low-reliability effect is therefore dominated by the overall strong fixed effects. In contrast, the bottom panels show a patient with a similar trajectory but with more data-points, indicating high reliability. This allows the unique random effect to be modeled in

line with the respective raw scores. This procedure is a practical way of weeding out single unreliable measures. This feature should be approached with caution when evaluating idiosyncratic developments where patients supply few data-points.

3.4.3 Missing Data. Assessing naturalistic psychotherapy data usually entails managing the issue of missing data. This is close to inevitable as patients might drop out of therapy, questionnaires can be misunderstood, get overlooked, or get lost in the mail. Missing data can be subdivided into categories that relate to the causal mechanism that produced it (Little & Rubin, 2019). The best-case scenario, from a statistical inference point of view, is when Missing data arise Completely At Random (MCAR). Data can be said to be MCAR when the missing data is completely unrelated to any variable that might influence the research questions. Random letters getting lost in the mail is an example of MCAR. Unfortunately, the MCAR assumption is seldom appropriate for psychotherapy data as the processes of dropping out of treatment or failing to deliver a questionnaire could be related to events that are related to the research question at hand.

When assessing missing data that cannot be assumed to missing completely at random there are two main approaches. If the missing data can be accounted for by another observed variable, or Missing At Random (MAR), then it can be managed by using statistical inference. Specifically, if the likelihood of missing is perfectly captured in one or more observed variables, then these can be used to provide unbiased estimates of the missing data. An example of this is the scenario were patients only drop out of treatment due to low alliance with their therapist and this level of alliance is measured by the researchers. If the missing data cannot be deemed either MCAR or MAR, then it can be said to be Missing Not At Random (MNAR). This is the highly problematic issue facing researchers who do not believe that their missing data is random and cannot use other measured variables to infer missing values. Data is MNAR when the missing values are related to the cause of missingness. One

example of this is if a patient in a psychotherapy trial drops out because of paranoid delusions towards the therapist and the missing values are related to the measurement of distrust of others.

Depending on the classification of missingness, missing data can be both harmless and precarious for statistical inference in longitudinal psychotherapy trials (Enders, 2011). There are a number of different approaches to managing missing data that attempt to minimize its impact. Firstly, the simplest procedure is simply to ignore it or exclude all patients with missing data present, called listwise deletion. This is an adequate approach when data can be safely assumed to be MCAR. Another option is to utilize all available data to estimate statistical properties that do not require several measures, e.g. means and variances, but to delete values that do not have a corresponding later measurement, e.g. deleting any pretreatment values for patients missing posttreatment values. This approach is called pairwise deletion. The benefit of this approach is that more of the data is utilized which increase statistical power. The downside is that it can lead to biased estimates of standard errors. Both listwise and pairwise deletion is heavily utilized in longitudinal psychological research and has been labeled as a bad practice for handling missing data (Jeličić et al., 2009). An alternative approach is to impute missing values using a conservative measure such as an overall mean or using the Last Observation Carried Forward (LOCF) in a longitudinal design. Both mean and LOCF imputation have been shown to lead to bias and a reduction in precision (Hamer & Simpson, 2009). More appropriate tools for managing missing data include multiple imputation and likelihood-based approaches (Little & Rubin, 2019). Multiple imputation seeks to utilize all observed values in order to impute a missing value. This is usually accomplished by an iterative process where several estimates are attempted utilizing a Bayesian probability model. Imputed values are selected by converging on parameters that produce the highest likelihood. This feature is particularly beneficial in the case of

psychotherapy trials such as the NMSPOP. As the NMSPOP gathers a great deal of background and treatment data, this can be used in order to make educated imputations, in contrast to investigations that focus on a particular issue with few measures.

One of the main befits of utilizing multilevel models in longitudinal research is the ability to handle missing data. Specifically, MLM's allow for nonparallel waves of missing data across individuals (Tasca & Gallop, 2009). This is because MLM's utilize every available data point in order to estimate variance components and fixed effect. Unlike multiple imputation, data is not imputed but no data is disregarded to produce parameter estimates. Although MLM's are by no means a panacea for managing missing data in longitudinal psychotherapy research, they have been shown to outperform deletion methods and simple imputation procedures (Hamer & Simpson, 2009).

3.5 Cohens Kappa

As with any measure, the diagnostic assessments are not free bias. Research has demonstrated that inter-rater reliability can be an issue when measuring diagnostic status with the SCID I and II (Lobbestael et al., 2011). Cohens Kappa was used in order to assess inter-rater reliability of the raters. In contrast to a simple percentage agreement calculation, Cohens Kappa incorporates the base rate agreement. It ranges from 1 (perfect agreement) to 0 (agreement no better than random). A sample of 40 SCID I and 20 SCID II interviews were selected at random to assess inter-rater reliability. These were blindly double coded by independent raters in order to assess inter-rater reliability.

3.6 Ethical Considerations

A central goal of psychotherapy research is to explore how patients respond to psychotherapy and which conditions play a part in differentiating the response to treatment. This entails a host of normative evaluations that can either be managed through implicit ethical laden actions or through explicit ethical consideration (Berg & Slaattelid, 2017).

Examples of such practice are the act of defining what constitutes as good or bad outcomes, establishing thresholds that differentiate patient responses, choosing the goal of research, and setting statistical margins of errors.

Patient participation in outcome research also presents itself with a set of ethical dilemmas. The research presented in this thesis is closely tied to a healing practice that each patient requires, which is typically delivered in a hospital setting. Patients rarely actively seek out participation in a research program, but rather seek the alleviation from a mental illness. One central goal of this project is to investigate a patient cohort that is similar to that which can be seen in a regular outpatient treatment clinic. This entails intervening in the status quo of healthcare delivery to offer participation in a research program. In addition to a set of practical challenges, this presents a set of unique ethical considerations. Although patients were explicitly informed that participation was voluntary and that they could disengage from the project at any time, there still exists a possibility of unwanted and unconscious coercion. This is of particular importance as patients might feel obliged to participate in order to secure their treatment. In addition, some patients in a traditional healthcare setting suffer from levels of mental illness which can entail a reduction of the capacity to decline a research invitation. For instance, a patient with a dependent personality disorder may fear rejection if he or she declines an invitation to participate in research.

The challenge of research collaboration with a patient cohort characterized by moderate to severe mental illness can be sidestepped by instead focusing on a cohort that is less severely afflicted with mental illness and is not engaged in, or served by, routine healthcare services. This downside of this strategy is that this research cannot be generalized to the more severe cohorts. This entails that the benefits of research, such as increased understanding, development of clinical tools, more efficient treatment, and refinement of practice, might not reach the severely afflicted cohort (Dunn, 2016). This is exemplified by

the psychotherapy dosage literature. A wealth of evidence suggests that the required dosage of psychotherapy depends on the patients' level of psychopathology (Baldwin et al., 2009). In a recent systematic review, Robinson et al. (2019) concluded that 95 % of patients from psychotherapy dosage investigations come from university counseling centers. Without investigations of severely afflicted patient cohorts, the estimates of required amounts of psychotherapy will be heavily biased toward low treatment dosages.

3.6.1 Registration Prior to Analysis. The last ten years have seen a major shift in psychology towards open science practices that include pre-registration of hypothesis, openness of analysis tools, and sharing of data and materials used in analysis. This shift is the result of a series of important psychology findings that have been shown not to replicate when examined by other researchers (Maxwell et al., 2015). A number of problematic practices have been pointed out as potential sources of bias such as unreliable measures, low powered studies, p-hacking, publishing bias, and the practice of Hypothesizing After the Results are Known (HARKing). Some authors have pointed out that many of the proposed issues can be understood as unconscious or unwise research practices, in contrast to deliberate scientific fraud (Chambers, 2017). The work presented in this thesis has attempted to ameliorate some of these pitfalls by registering hypotheses and planned analyses before carrying them out. The pre-analysis registrations, available in the appendix, serve several goals. By simulating a dataset with similar qualities, the researcher can establish and experiment with different analysis tools without the risk of getting influenced by the results they produce. The correct analysis tool can be operationalized in detail using transparent analysis scripts. Also, formulating specific hypotheses before analysis negates potential HARKing. The author of this thesis believes that this approach can be utilized for reanalysis of psychotherapy research data when no pre-registration has been performed, in order to fully utilize psychotherapy

research data. This is particularly relevant as psychotherapy investigations are both expensive and possibly burdensome for patients and therapists.

4. Results

Paper I: Effectiveness of Open-Ended Psychotherapy under Clinically Representative Conditions

The first paper demonstrated that, when treated with an open-ended format, the length of therapies for a heterogeneous patient population was equally variable with a mean number of sessions of 51.3, a standard deviation of 58.9, and a median of 35. When comparing pre- and posttreatment levels of overall symptoms and interpersonal problems we found large (.85) and moderate (.57) effect sizes respectively. These effects were stable when reassessed through the two-and-a-half-year follow-up. A large proportion of patients recovered (GSI: 38 %, IIP-Global: 23 %), as defined by demonstrating both a statistically robust change as well as crossing a theoretical cutoff between an abnormal and normal population. The patient that did not recover either experienced improvement (GSI: 31 %, IIP-Global 12 %), no change (GSI: 29 %, IIP-Global: 63 %) or deterioration (GSI: 1.6 %, IIP-Global 2.4 %). Patients also reported a significant reduction of both DSM – 4 Axis I (clinical disorders) and II (personality disorders) diagnoses. At pretreatment, 87 % reported one or more clinical disorders, followed by 40 % at posttreatment and 38 % at a two-and-a-half-year follow-up. Similarly, 54 % of the sample experienced one or more personality disorders at pretreatment, dropping to 28 % at posttreatment and 22 % at follow-up. In contrast, patients did not experience a statistically significant improvement in occupational functioning when comparing both posttreatment and follow-up with pretreatment measures. The substantial variability in treatment length, high heterogeneity in patient characteristics and diagnoses, as well as overall severity of the sample, entail a high degree of ecological validity that few other psychotherapy studies provide.

Paper II: Comparing the magnitude of improvement for patients with and without personality disorders during open-ended psychotherapy

In paper II, we found that patients with a personality were more severely afflicted in all measures of psychopathology across all timepoints. This includes psychiatric symptoms, interpersonal problems, and frequency of Axis I clinical disorders, as identified by the SCID interview. We found that patients with a personality disorder experienced equal symptomatic improvement and greater interpersonal improvement compared to patients without. The no personality group demonstrated relatively minor interpersonal problems at pretreatment. Both patients with and without a personality disorder lost their Axis I clinical disorder at the same rate throughout treatment. However, we found that patients with a personality disorder at pretreatment were roughly twice as likely to regain a previously lost clinical disorder in the follow-up phase. In terms of self-evaluated psychiatric symptoms and interpersonal problems, both groups demonstrated enduring improvements when self-assessed at a two-and-half-year follow-up. We also found that the degree of personality pathology was positively related to the magnitude of change. Patients with more severe personality problems, as measured by frequency of qualified SCID II items, demonstrated greater gains in the open-ended treatment format. Using marginal models, we found that a patient qualifying for zero SCID II items, indicating no personality problems, showed a predicted pre-post within Cohen's *d* of .53 on the GSI while a patient qualifying for 30 items showed a predicted Cohen's *d* of 1.25. However, despite making substantial gains during therapy, the patients with a personality disorder were still afflicted with a psychopathology posttreatment. The patients with a personality disorder ended up at approximately the same level of psychopathology as the no personality group experienced pretreatment.

Paper III: Patients with different levels of psychopathology have different psychotherapeutic needs

In paper III, we demonstrate that patients that are afflicted with severe psychopathology requires more sessions in therapy, but demonstrate larger magnitudes of improvement, as measure by overall psychiatric symptoms and interpersonal problems. In contrast, patients with milder conditions demonstrated a faster treatment response that was smaller in magnitude. The overall best-fitting model of change during psychotherapy was found the be a linear effect model where the benefit of time in therapy increases linearly in interaction with a dose variable. This stands in contrast to previous research which suggests that the potency of therapy rapidly diminishes throughout treatment.

The largest magnitudes of change were found in patients who attended very long treatments, although exact specifications could not be made due to the relatively few patients undergoing very long treatments. When defining outcomes in terms of level of functioning at posttreatment, the patients with mild conditions had superior outcomes. Conversely, when defining patient outcomes in terms of total improvement achieved, patients with more severe conditions demonstrated superior outcomes. We also assessed improvements in terms of the probability of recovering during treatment using survival analysis. This analysis demonstrated a pattern of gradual diminishing therapy potency as reported by other similar investigations. In contrast to previous work, our results reveal that improvements continue through long treatments. We found a median survival time of 57 sessions which is substantially larger than previous estimates which range between 4 – 20 sessions. The paper argues that the majority of psychotherapy dosage recommendations are based on patient cohorts that are disparate from what is typically seen in a Norwegian routine outpatient psychiatric clinic. Lastly, the results indicate that some very long treatment failed to produce a positive change indicating that open-ended and flexible psychotherapy is not a magic bullet for achieving recovery.

5. Discussion

5.1 General Discussion of Findings

Overall, the findings from the NMSPOP adds to the psychotherapy effectiveness literature by including longer and more flexible treatments with a diverse patient cohort. The treatments delivered were flexible in the sense that they were not directed by a particular research protocol or treatment manual, and that the number of sessions delivered, and the frequency of those sessions was tailored according to clinicians' patients' collaborative evaluation of the needs of each patient. This arguably allows for greater understanding and generalization to the patients that are treated in traditional outpatient clinic setting which often includes high variability in the patient cohort, as well as therapists who do not adhere to a particular protocol (Gkeredakis et al., 2011; Kazdin, 2017).

The results demonstrated that a substantial majority of patients experience considerable mental health improvements that are usually maintained throughout at least a two-and-a-half-year follow-up period. We found that patients have highly variable psychotherapeutic needs with correspondingly high variability in the psychotherapy dosages received. This variation seems to be closely linked to the patient level of overall psychopathology at the onset of treatment.

This thesis concludes that the severity of psychopathology is negatively associated with the rate of change during treatment, while positively associated with the overall magnitude of change when treatment is personalized in terms of intensity and duration. Patients with more severe conditions require larger therapy dosages in order to achieve amelioration, but at the same time also show the greatest potential for overall change when treated with such a format. This pattern is also reflected when comparing patients with and without a PD. Patients with more severe characterological problems require more therapy but achieve the greatest magnitudes of change. The results from this thesis add to a growing

literature on psychotherapy outcomes in a naturalistic setting. The NMSPOP-project is arguably rare in the sense that it employs therapists who are under no theoretical restrictions, treating patients who are diverse and representative for a varied outpatient clinic, and utilize a rigorous research methodology, including systematic, observer-rated diagnostic assessments. In many ways, these results reveal a hopeful prospect of amelioration and recovery for patients across the spectrum of psychopathology severity.

5.2 Treatment and Time Effects

The patients treated in the NMSPOP trial experience a range of influences on their mental health, in addition to receiving psychotherapy. They could win the lottery, make a new friend, or marry their lover. Conversely, they could also experience bereavement, lose their job, or become injured. Life events such as these have a strong link to the development and maintenance of psychopathology (Eisenbarth et al., 2019). Researchers examining psychopathology should therefore not assume it to be a static concept, but rather a highly sensitive and fluctuating phenomenon (Lahey et al., 2017). A researcher interested in group-level changes across time treats this individual variance as stochastic noise which is to be parsed out using tools aimed at central tendencies. If the ups and downs of life equal out, then the effect of time should be zero. If patients who undergo psychotherapy demonstrate a better overall development across time, as measured by a statistic of central tendency, compared with people that do not receive therapy, then the researcher might conclude that the positive effect is caused by the treatment.

The results of this thesis are all observational, in contrast to an experimental design where study variables can be controlled and manipulated. The improvements seen in the NMSPOP-cohort can be attributed to a mix of two main causal influences, namely the treatment received and an opaque and spontaneous benefit-over-time-effect. Research on the development of untreated psychopathology is sparse, but it does suggest that mental illness

can recede without intervention (Meares et al., 1999; van Beljouw et al., 2010; Whiteford et al., 2013; Zanarini et al., 2010). Although this spontaneous recovery-effect is well-documented, the same evidence reveals that the effect is meager in comparison to the treatment effects of psychotherapy. For instance, Bateman & Fonagy (2008) compared patients treated with MBT with a control condition who were offered treatment-as-usual, which consisted of little to no psychotherapy. An eight-year follow-up revealed that 13 % of the MBT patients could be classified as suffering from BPD, compared with 87 % in the treatment-as-usual group. Also, a meta-analysis by Cuijpers et al. (2014) comparing psychotherapy for adult depression with placebo medication, found a Hedges *g* effect size of 0.25 in favor of psychotherapy. Similarly, there is an abundance of research indicating that psychotherapy outperforms a wait-list control condition (Dragioti et al., 2017). The wait-list control is the most utilized control for the spontaneous recovery effect. However, it should be noted that the use of wait-list control has recently been shown to be deleterious for many patients (Patterson et al., 2016; Steinert, Stadter, et al., 2017), indicating that it serves as a poor representation of the natural course of mental illness (Cuijpers, Karyotaki, et al., 2019).

In this thesis, it is speculated that the spontaneous benefit of time is small due to the high levels of chronicity and severity of psychopathology reported by a large proportion of the NMSPOP cohort. However, the goal of this thesis is not to establish whether there is an effect of open-ended psychotherapy, as such, as this is arguably established (L. F. Campbell et al., 2013). Rather, the aim is to evaluate the changes seen in patients and to investigate hypotheses related to theories of psychotherapeutic change. The results from this thesis can also be utilized as a benchmark for therapists and policymakers that require data on what outcomes can be expected when offering open-ended treatments to a wide array of patients with different levels of psychopathology.

5.3 Treatment Format and Patient Outcomes

A central issue throughout this thesis is the relationship between a particular psychotherapy treatment format and how this might influence patient outcomes. The underlying proposition is that variations in treatment formats cause systematic variation in response to the treatment. If, for instance, a highly delimited and short treatment fails to treat patients with more severe psychopathology because they require a more flexible or a longer treatment format, then this is an example of such a relationship.

Unfortunately, this causal claim cannot be assessed using our current methodology. This is because the causal role of treatment format as a moderator for therapy outcomes is confounded by a host of other variables. Without a specified control condition, we cannot conclude that our results are due to a particular aspect of the treatment delivered. Nonetheless, the results from this thesis are consistent with the assertion that patients have highly variable psychotherapeutic needs, as indicated by the great variability in dosage received. The temporal dimension of change is also consistent with the claim that psychotherapeutic interventions requires more time in treatment for more severely afflicted patients, given the assumption that time spent in treatment causes the improvements. It should also be noted that the results indicate no relationship between treatment length and changes during follow-up. This implies that even the short treatments in our study were associated with lasting effects. This observation should be viewed in light of the fact that the length of treatment was a choice made by patients and therapists in collaboration, in contrast to a fixed dosage scenario. Another interpretation consistent with our results is that some patients and therapists are correctly convinced that they do not require long treatments.

The relationship between treatment format, patient characteristics, and patient outcomes are central to further development of psychotherapeutic interventions. This relationship could in principle be investigated by experimental research that can tease out

causal relationships. Such a study could randomize patients into different levels of treatment dosages as well as provide an open-ended condition. The results of this study could be analyzed in terms of both the main effect of the experimental condition, but also the interaction effect with patient severity levels and comorbidity. However, an experimental dosage study of naturalistic treatment raises ethical concerns as patients with severe levels of psychopathology might require a particular treatment length. Experimentally manipulating the dosage might entail disregarding the needs of the patient to accommodate the research agenda. This ethical dimension of dosage may explain why the dose-response psychotherapy literature is almost entirely comprised of university counseling clinic samples, where treatment is arguably less of a medical necessity.

An important distinction should be made between outcomes defined as a posttreatment level of functioning, and outcomes defined as the total improvement demonstrated by the patient. As noted earlier, a patient can experience significant positive improvement, but still experience a relatively high level of psychopathology posttreatment. This is illustrated in both paper II and III where the patients with the most severe psychopathology demonstrated the largest overall gains. When comparing the severely afflicted cohort with the mild, the latter had less improvement but better posttreatment functioning. Brown et al. (2001) found a similar pattern when analyzing a large dataset of patients treated in managed care organizations. The authors concluded that although the most severely afflicted patients demonstrated the overall largest overall gains, they seldom improved to the point of recovery. Similarly, Hansen et al. (2006) reported that greater pretreatment severity was associated with greater improvement when assessing naturalistic-setting data. Studies of long-term psychotherapy treatments indicate a similar pattern (Leichsenring & Rabung, 2011; Lorentzen & Høglend, 2004).

The thesis findings contradict research which indicates that severity and/or comorbidity is negatively associated with improvements. In the Newman et al. (2006) review of the anxiety treatments, they concluded that comorbid depression, personality disorder, and substance abuse all negatively influenced outcomes. Beutler et al. (2006) found that the majority of investigations on treatment for depression and dysphoria indicated that comorbid personality disorder negatively influenced outcomes. After summarizing this research literature, Bohart & Greaves (2013) tentatively concluded that severely afflicted patients and patients with comorbidity may need prolonged and flexible treatment in order to ascertain equivalent outcomes. Similarly, McAleavey's (2019) meta-analysis of psychotherapy effectiveness studies concludes that highly distressed patients were inadequately treated and that this inadequacy was possibly related to the brevity of treatment provided. The results from this thesis are compatible with the assertion that severely afflicted patients require a far greater treatment length compared with mild and moderately afflicted patients.

5.4 Psychotherapy Efficacy and Effectiveness

The majority of the outcome literature utilizes a specific treatment program for patients within a particular diagnostic group. Some examples are cognitive behavioral therapy for panic disorder (Stewart & Chambless, 2009), mentalization-based therapy for borderline personality disorder (Sharp & Kalpakci, 2015), or interpersonal therapy for depression (de Mello et al., 2005). In this area of research, large variation in the severity of mental illness and comorbidity are possible confounders that can lower a study's internal validity. The balancing act between experimental control and ecological validity is at the heart the debate surrounding the implementation of evidence-based psychotherapy (Nathan et al., 2000; Westen et al., 2004). Low ecological validity means that the results of efficacy research done in an RCT-setting might not generalize to real-world naturalistic settings where psychotherapy is commonly delivered.

What constitutes naturalistic therapy is a moving target with large differences across countries and health regions. There is arguably a global trend of decreasing the gap between the RCT setting and naturalistic treatments by establishing standardized treatment protocols and patient evaluations. The Improving Access to Psychological Therapies initiative (Clark, 2011) in the UK and Norway (Knapstad et al., 2018) are examples of this development. The move towards standardized treatment protocols may be related to the finding that outcomes from RCT's often outperform naturalistic outcomes (McAleavey et al., 2019). Similarly, naturalistic treatments with conditions that mimic the RCT setting has been shown to produce better outcomes compared with interventions that do not (Shadish et al., 2000). Results from meta-analyses indicate that certain study features are associated with improved outcomes, particularly increased therapy dosage and the utilization of specific psychopathology outcome measures (e.g. Panic Disorder questionnaire). These are in contrast to general outcome measures (e.g. SCL-90-R). The work presented in this thesis, as well as the majority of psychotherapy effectiveness investigations, utilizes general outcomes measure because they allow for comparison across different diagnostic categories. A downside is that potentially clinically important idiosyncratic developments, specific to the patient's psychopathology, are not measured.

Pragmatic studies of psychotherapy effectiveness across a wide range of different diagnoses and severity levels will tend to produce lower effect sizes compared with specialized interventions for a specific mental disorder. This is because most operationalizations of effect size, such as Cohen's *d,* is determined by relating the change in raw score to the variability found in the measure. If a study population has low variability, then Cohens *d* is higher. This can be illustrated by comparing the NMSPOP sample to a more restricted RCT sample. Puolakanaho et al. (2020) delivered an Acceptance and Commitment

intervention to alleviate burnout symptoms. They report a pretreatment Global Severity Index Standard Deviation of .43, compared with .61 in the NMSPOP cohort.

Table 1 demonstrates how two standard deviations give way to different effect sizes given $d = \frac{\text{Raw Change}}{\text{Standard Deviation}}$.

Table 1: Cohens *d* from Two Standard Deviations

Raw Change	SD = .43	SD = .61
0.20	.47	.33
0.40	.93	.66
0.60	1.4	.98

This methodological quirk can have dramatic effects if comparisons between studies rely on treatment effects as defined by Cohens *d* or similar standardized effect size measures that are highly sensitive to between-patient variance. This observation should serve as a caution for indiscrete comparison of effect size measures across different patient populations. A more prudent approach would be to assess raw change when making this comparison, although this requires intimate knowledge regarding the specific outcome questionnaire.

A similar issue arises in the measurements that are utilized in effectiveness trials in comparison with specialized interventions. In order to make between-patients comparison, effectiveness trials rely on global estimates of mental illness, such as the GSI or IIP Global. As previously shown, these are less sensitive to change compared with measures of a particular mental disorder, e.g. measuring the tendency to panic while treating exclusively patients with a panic disorder.

5.4.1 Comparing American University Counselling Center & NHS cohorts with NMSPOP

The largest source of outcome variation in any psychotherapy investigations is, arguably, the client (Norcross & Lambert, 2011). It therefore follows that the selection criteria utilized in research have great power to shape the results from any psychotherapy outcome

investigation. As previously shown, both the dosage literature and the naturalistic psychotherapy outcome literature overwhelmingly originate from data collected at American university counseling centers. These are clinics that are dedicated to providing mental healthcare to students on select campuses. The data sources that are not from American university counseling centers are usually either treatment data originating from the United Kingdom's National Health Service (NHS) or data from specialized treatment RCT-interventions. There is a tacit claim throughout all three papers presented in this thesis that the NMSPOP cohort is more severely afflicted with psychopathology compared to American university counseling clinics and investigations of NHS-based outcomes.

Unfortunately, there is no straightforward way of comparing the severity of these cohorts as they measure patients with different outcomes measures. Also, naturalistic investigations seldom offer more than a single self-assessed measure of mental health which limits comparison. A single outcome measure does not allow for a wide-specter evaluation of mental health to the same extent as an evaluation which includes diagnostic assessments. Unfortunately, diagnostic evaluations are rarely administered due to their high cost and practical challenges. A Z score can be used as a coarse value of comparison between studies. The Z score in table 2 is calculated for GSI, OQ-45, OQ-30, BHM-20, CCAPS, and CORE-OM utilizing non-clinical means and standard deviation from Derogatis & Unger (2010), Lambert et al. (1996), Ellsworth et al. (2006), Green et al. (2003), Lockard et al. (2012) and Barkham et al. (2005), respectively. The Z score represents the number of standard deviations the particular study cohort, at pretreatment, is from the nonclinical mean score.

$$Mean\ Nonclinical\ Cohort +/-(Z\ Score * SD\ Nonclinical) = Mean\ Clinical\ Cohort$$

Table 2 includes studies that assess the effectiveness of psychotherapy in a routine-care setting, utilizing either OQ-30, OQ-45, SCL-90-R, BHM-20, CCAPS or CORE-OM. The search engine for this informal literature review was Google Scholar ®, utilizing the

following keywords: PSYCHOTHERAPY, OUTCOME, EFFECTIVENESS, NATURALISTIC, TREATMENT-AS-USUAL, DOSAGE, DOSE-RESPONSE. Studies were also sourced from recent literature reviews (McAleavey et al., 2019; Robinson et al., 2019).

Table 2: Studies Identified in the Literature Review

	N	Mean Sessions (SD)	Mean Pretreatment (Measure)	Z Score	Median Time-To-Recovery	Setting
NMSPOP	**370**	**51.3 (58.9)**	**1.28 (GSI)**	**2.48**	**57**	**Norway Outpatient Care**
Asay et al. (2002)	29	N/A	80.6 (OQ-45)	1.8	42-54	U.S. Private Practice
Carr et al. (2017)	132	N/A	N/A (OQ-45)	N/A	7 - 19	U.S. University Clinics
Draper et al. (2002)	1698	3.3 (2.4)	N/A (OQ-45)	N/A	4- 10	42 University Clinics 9 Private Practices All U.S.
Erekson et al. (2015)	21488	5.8 (4.2)	68 (OQ-45)	1.1	8	U.S. University Clinics
Falkenström et al. (2016)	924	Primary: 6 (3.1) Psychiatric: 9.1 (5.3)	1.7 (CORE-OM)	1.6	8 - 13	Primary Care (n = 640) Psychiatric Outpatient Care (n = 284 All Sweden
Harnett et al. (2010)	125	9.5 (Median: 8)	79 (OQ-45)	1.7	10-14	Australia University Clinic
Kopta et al. (1994)	854	N/A (Median: 15)	1.52 (GSI)	3.1	58	U.S. Mental Health Services
Kopta et al. (2014)	13803	5.7 (4.8)	2.56 (BHM)	1.5	7 - 10	U.S. University Clinics
Lutz et al. (2005)	203	< 16	1.33 (CORE-OM)	0.97	N/A	U.K. NHS Primary Care
McAleavey et al. (2019)	9895	6.49 (3.9)	1.37 (CCAPS)[a]	0.8	N/A	U.S. University Clinics
Minami et al. (2008)	12743	4.2 (5.2)	62 (OQ-30)	2.1	N/A	U.S. Managed Care
Owen et al. (2015)	10854	9.41 (4.39)	2.37 (BHM)	1.9	11	U.S. University Clinics
Owen et al. (2016)	13664	9.04 (5.7)	2.25 (BHM)[b]	2.2	N/A	U.S. University Clinics
Puschner et al. (2008)	256	22.2 (18.8)	1.02 (GSI)	1.8	N/A	Germany Outpatient Clinic
Reese et al. (2011)	1207	7.8 (7.9)	75 (OQ-45)	1.5	N/A	U.S. University Clinics
Saxon & Barkham (2012)	10786	5.9 (3.0)	1.75 (CORE-OM)	1.7	N/A	U.K. NHS Primary Care
Snell et al. (2001)	158	6.4 (5.1)	N/A (OQ-45)	N/A	14-16	U.S. University Clinics
Stiles et al. (2008)	9703	N/A (Median: 6)	19 (CORE-OM)	1.3	N/A	U.K. NHS Primary Care
Stiles et al. (2015)	26430	8.3 (9.5)	19 (CORE-OM)	1.3	N/A	14 University Clinics 6 Primary Care Centers 8 Secondary Care Centers 10 Tertiary Care Centers 8 Workplace Interventions 2 Private Practice Centers All U.S.
Stulz et al. (2013)	6375	8 (7.2)	2.24 (BHM)	2.1	N/A	26 University Clinics 4 Primary Care Centers 2 Private Hospitals All U.S.
Watzke et al. (2012)	147	18 – 25 (N/A)	1.17 (GSI)	2.2	N/A	Germany Outpatient Clinic
Wolgast et al. (2003)	788	5.87 (Median: 4)	71 (OQ-45)	1.3	14-16	U.S. University Clinics

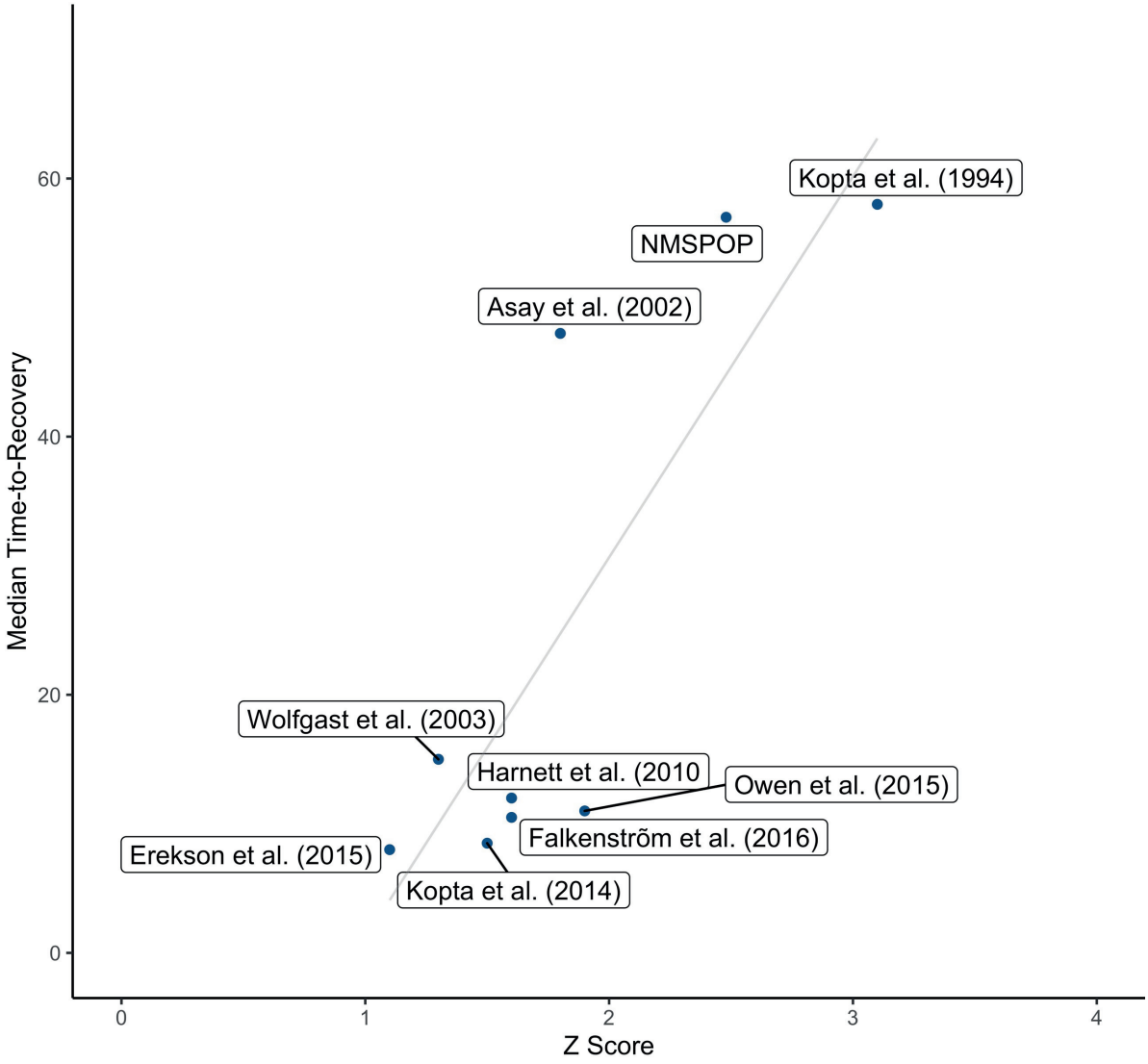
Derogatis & Unger (2010) Non-Clinical Mean (SD) SCL-90-R GSI = 0.34 (0.38)
Barkham et al. (2005) Non-Clinical Mean (SD) CORE-OM Total Score: 0.76 (SD = 0.59)
Lambert et al. (1996) Non-Clinical Mean (SD) OQ-45 Total Score: 48.2 (18.2)
Ellsworth et al. (2006) Non-Clinical Mean (SD) OQ-30 Total Score: 31.5 (14.22)
Green et al. (2003) Non-Clinical Mean (SD) BHM Global Mental Health: 3.29 (0.47)
Lockard et al. (2012) Non-Clinical Mean (SD) CCAPS: 0.78 (0.77)
[a] Mean Score across all six sub-measures
[b] Total Score not reported. Only Symptomatic Distress used

This comparison should be evaluated cautiously as it is an imperfect comparison that is very sensitive to sample variations, as well as variations in nonclinical comparison cohorts. A more accurate appraisal would require treatment samples which complete the same outcome measure. However, this investigation offers some evidence for the claim that the NMSPOP is more severely afflicted by mental illness when compared to studies originating from both American university counseling centers and investigations of NHS treatment data. Both

NMSPOP and Kopta et al. (1994) stands out with more severe Z scores and a high median time-to-recovery. Kopta et al. (1994) also report survival estimates for sub-measures based on subdividing the SCL-90-R items aimed at different types of psychopathology. Patients with "acute" symptoms required five, those with "chronic" symptoms required 14, and those with "characterological" symptoms required 104 sessions for 50% of patients to achieve clinically significant change. This observation demonstrates the problematic conceptualization of the drug metaphor as highlighted by Stiles & Shapiro (1989). Reducing the complexities of patient improvement to a single mean or sum score entails aggregating a complex and varied set of underlying improvement processes.

As demonstrated in figure 7, there seems to be a relationship between intake severity, as measured by the Z score, and median time-to-recovery. This finding provides between-study support for the claim that patients afflicted with more severe mental illness require longer treatments in order to recover.

Figure 7: Relationship between Time-to-Recovery and Cohort Severity



Arguably the most prominent finding from this review is the very large discrepancy between mean treatment length of the NMSPOP and most other investigations in routine-care. It demonstrated that an open-ended treatment policy can produce treatments that are very long, compared to more restricted policies. The NMSPOP also stand out from the majority of included studies with a high proportion of patients with a personality disorder. Asay et al. (2002) also reported high median time-to-recovery and a relatively low Z score. This study was based on a psychodynamic clinic investigating a single therapist treating a patient cohort severely afflicted with mental illness. Similar to the NMSPOP, a high proportion of patients were diagnosed with a personality disorder (66 %) and no strict session restrictions were

enforced. As shown in paper I and II, patient with a personality disorder require far longer therapies but also has the capacity for considerable magnitudes of change. The large proportion of patients with a personality disorder can also explain the very high median time-to-recovery.

In summary, the available outcomes for investigations of routine psychotherapy and psychotherapy dosage are comprised of studies that utilize a less severe patient cohort when compared to the NMSPOP trial. There is evidence from within and across studies suggesting that patients with more severe psychopathology require longer treatments to achieve recovery. It should be noted that this conclusion rests on the causal claim that it is the therapy, and not some other process, that causes the improvement. The potential curative effect of time is a confounder that is not controlled for in any of these studies. This issue is actualized by investigations that track patients for long treatment spanning several months or years.

5.5 Do Patients with a Personality Disorder Require Specialized Treatment?

There is an ongoing debate regarding the superiority of specialized interventions for personality disorder when compared to treatment-as-usual. There is a limited amount of evidence that suggests that specialized treatments produce better outcomes (Cristea et al., 2017; Oud et al., 2018). However, the superiority is modest and statistically unstable. Also, the term *treatment-as-usual* is largely opaque and unexplored in many of these investigations. Treatment-as-usual might involve an alternative mode of psychotherapy or no psychotherapy at all. In many studies, treatment-as-usual is synonymous with psychopharmacological interventions with a limited follow-up from a general physician (Widiger, 2012). There are a few studies that compare specialized interventions for borderline personality disorder with a known treatment-as-usual condition. Clarkin et al. (2007) compared transference-focused therapy, dialectical behavioral therapy, and supportive dynamic therapy. The latter was assessed to be similar to treatment-as-usual in the context of outpatient clinical care. No

treatment was found the be superior. McMain et al. (2009) randomized patients into either dialectic behavioral therapy or general psychiatric management. The latter category was evaluated to be in line with APA's guidelines for the treatment of borderline personality disorder. The results demonstrated no differences between the two conditions. Similarly, Bateman & Fonagy (2008) randomized patients to either mentalization-based therapy or structured clinical management and found no differences in outcome. Jørgensen et al. (2014) randomized patients to either a high dosage mentalization based intervention or group-based support interventions. No difference was found between the two conditions at posttreatment and 18 months after treatment. This result is particularly noteworthy as the former group received a far more intense treatment and treated individually, in contrast to the group support conditions. Therapists in both conditions were highly experienced, utilizing either psychodynamic or psychoanalytical techniques. Both treatment conditions were monitored and closely supervised.

In contrast to studies purely focusing on borderline personality disorder, Antonsen et al. (2014) recruited patients with a personality disorder in general. The authors excluded patients with antisocial and schizotypal personality disorders. The majority of the included patients (N = 113) suffered from either borderline (46 %), avoidant (41 %), unspecified (21 %), obsessive-compulsive (9 %) or dependent (7 %) personality disorder. Patients were randomized into either a step-down intensive treatment or a traditional outpatient treatment. The former consisted of an initial 18-week inpatient hospital treatment with 3-4 therapy sessions each week, followed by outpatient care which consisted of weekly group- and individual psychotherapy for several years. Traditional outpatient treatment consisted of individual psychotherapy delivered by therapists in private practice, instructed to provide their usual practice. No restrictions or guidelines were put on treatment lengths. This group experienced a mean treatment duration of 56 sessions (SD = 56). Both treatment conditions

showed significant improvements, but no differences were found between the two groups at posttreatment, nor at a three- and six-year follow-up.

In conclusion, specialized interventions seem to outperform treatment-as-usual in studies where there is little control or monitoring of what therapy is delivered. However, specialized treatments have not been shown to outperform treatment-as-usual when the latter condition delivers individual psychotherapy. In their review, Livesley and Larstone (2018) conclude that structured psychotherapy interventions, performed by professionals under supervision, cannot be said to be inferior to outcomes produces by specialized treatments for borderline personality disorders. The findings from this thesis are consistent with the assertion that treatment-as-usual, defined as flexible outpatient psychotherapy, produces substantial and stable improvements for patients suffering from a personality disorder. The NMSPOP cohort is unique in its inclusion of a very broad spectrum of psychopathology. Another conclusion is that patients with more severe levels of personality problems demonstrate larger magnitudes of improvements in comparisons with less severely afflicted patients. This goes against an argument that the more severely afflicted patients require specialized interventions.

Overall, the results of this thesis are consistent with the hypothesis that the crucial ingredients of effective treatment of personality disorder are flexibility in treatment length and intensity, in contrast to specific treatment methodologies or interventions. It should also be noted that far from all patients experienced statistically reliable improvements or recovery. Some patients spent years in therapy without demonstrating a stabile positive response. This observation should serve as an impetus for the continued development of psychotherapeutic practice and research.

5.6 Fixed Dosage and the Good Enough Level of Functioning

The findings from this thesis are in line with a large body of empirical investigations which demonstrate that treatment length moderates rates of change (Baldwin et al., 2009;

Michael Barkham et al., 2006; Falkenström et al., 2016; Owen et al., 2016; Reese et al., 2011). This observation is a prediction of the good-enough level model of psychotherapy which posits that the psychotherapeutic processes differ across treatment lengths (Michael Barkham et al., 2006). This is in contrast to a fixed dosage perspective which predicts equal rates of change across all patients with different treatment lengths (Howard et al., 1986). These two models of psychotherapy are associated with distinct tacit assumptions regarding the nature of psychotherapy itself. The dose-response view utilizes a pharmaceutical analogy where psychotherapy treatment is framed within a decontextualized medicalized discourse (Jørgensen, 2019).

Hoffman (2009) describes how psychotherapists are faced with a double-edge sword of scientific legitimacy. On the one hand, there is the ambition of strengthening psychotherapy's position by utilizing research methodology borrowed from established medical science. On the other, psychotherapists are forced to recognize that key processes and outcomes cannot be reduced to simple operationalized concepts. The introduction of the dose-response relationship placed psychotherapeutic practice closer to medicine, thereby increasing its legitimacy in comparison with psychiatric medicine (Kopta, 2003). Other contemporaries critiqued this maneuver for methodological reductionism, arguing that several empirical findings falsify the drug metaphors tacit assumptions (Stiles & Shapiro, 1989). These include the observation that patients show idiosyncratic treatment trajectories, that active ingredients cannot be separated from the overall therapeutic context and that therapists do not passively deliver dosages of psychotherapy.

The good enough level model was proposed as an alternative to the dose-response view of psychotherapy and includes the psychotherapeutic processes as a key dimension for understanding psychotherapy dosages. Patient differ in psychotherapeutic needs and this variation is the foundation for different outcome trajectories. Barkham et al. (2006) argued

that the previously reported negatively accelerating treatment trajectory reflected a progressive ending of treatment by clients who had achieved a good enough level of improvement and not an effect of diminishing therapy potency.

The finding that treatment length moderates the rate of change has been resolutely replicated across a range of different conditions and could be regarded as one of the most robust findings from psychotherapy outcome research (Falkenström et al., 2016). These results have a normative component as they relate to the decision of limiting psychotherapy from a cost-benefit perspective. It can be argued that fixed dosages are a disservice to patients as their needs vary greatly. In a recent qualitative review, De Geest & Meganck (2019) conclude that time-limiting psychotherapy can have both a positive and negative influence on psychotherapeutic processes. It can promote a positive sense of urgency in both patient and therapist, focusing the treatment to make the most of it. Conversely, some therapists report that restrictive time-limits entail more forced direction from the therapist and that this leads to superficial therapy that does not stimulate the patient's therapeutic autonomy.

5.7 Evaluating Outcomes Continuously or Discretely

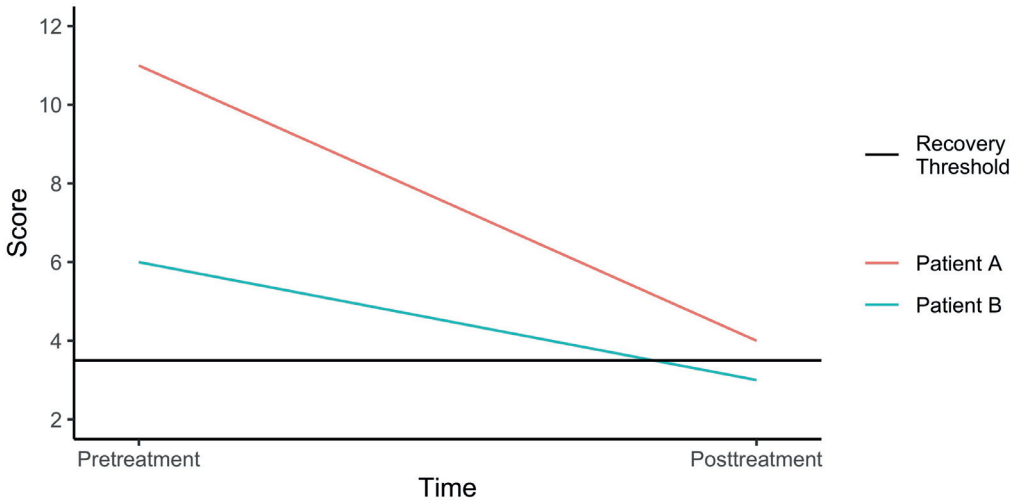
This thesis and the articles contained presents two complementary modes of assessing treatment outcomes. One option is to analyze improvements as a continuous process that corresponds to the changes seen in statistical models of outcome measures. These models either encompass improvements for the treatment group as a whole, assessed by a fixed effect using MLM terminology or incorporate patient-level improvements as random effects parameters. As previously shown, a MLM fitted with both a random intercept and random slope can be utilized to give an estimation of particular patients' trajectory as patient-level intercepts and slopes are sampled from the parameter estimates of the overall model. A continuous outcome measure has the benefit of revealing the total amount of change for either the group or the individual. In contrast, many research questions in the field of outcome

research are aimed at answering questions regarding the discrete event of patient recovery. This is an either/or framing that rests on the assumption that patient recovery, as a discrete event, is a meaningful construct. Psychotherapy researchers have utilized a variety of statistical tools for modeling recovery such as logistic and probit regression, Kaplan-Meyer analysis, and Cox proportional hazard analysis (Fitzmaurice et al., 2011). These tools have originally been devised to be used in settings where the discrete event in question is both salient and of critical importance, such as the failure of a mechanical component in mechanical industry, or the death of a patient in medicine (Andersen & Keiding, 2005).

The majority of recovery modeling in psychotherapy is done utilizing a variety of operationalization laid out by Jacobsen & Truax (1991). The authors suggest that recovery should be thought of as containing three key principles; Firstly, in order to be classified as recovered, a patient needs to present a pretreatment score that indicates that he or she is in a diseased state. Secondly, the improvement observed following treatment should exceed a level of improvement one should expect by chance variability, called reliable change. Thirdly, if normative data is available, then this should be used to create a cut-off where the patients' score is closer to the mean of the normal population compared with the mean of the afflicted population. Patients are usually classified as demonstrating clinically significant change when all three conditions are met. The discrete event of clinically significant change is then selected to be the point of recovery from the disease. The advantage of this operationalization is that it focuses on individual patient outcomes while providing a meaningful frame of reference. This is particularly helpful when researching a process that fluctuates and is expected to be present in a normal sample. A psychotherapeutic goal of reducing anxiety to a level of zero, for instance, is unwarranted as assessments of anxiety in the normal population is higher than zero.

Although it has some appealing qualities, modeling recovery with survival methods has several noteworthy limitations. Firstly, in order to pinpoint the recovery event with a high temporal resolution, measures need to be completed with high frequency, preferably every session or week. This feature makes it untenable to use several questionnaires or questionnaires with many items as this would require too much time. It is similarly ill-advised to have a diagnostic interview before every session. This means that investigations of dosage and psychotherapeutic recovery rest on brief, self-administered questionnaires. This can be a particular obstacle if the patient's psychopathology is related to, or reduces, insight into psychological functioning. Secondly, using a cut-off that demarcates a normal and pathological population places emphasis on the patient cohort that is closest to the recovery threshold and can completely negate large positive changes seen in patients with very high levels of pathology, as shown in figure 8.

Figure 8: Two Hypothetical Outcome Scenarios



Although patient A demonstrates the greatest magnitude of improvement, this dramatic effect is nullified due to not crossing the recovery threshold. In contrast, patient B shows less improvement but enough to be credited with both a statistically reliable change and crossing the recovery threshold. This effect can bias outcome research to favor treatments that are aimed at the less severely inflicted with psychopathology. This effect was clear when

comparing patients with and without a PD in paper II. A large portion of patients with a PD demonstrated large improvements but could not reach the clinical recovery threshold.

Measuring change and applying personalized interventions holds great promise for future researchers of both medical and psychotherapeutic outcomes (Fröhlich et al., 2018; Zilcha-Mano, 2019). However, the practice of applying recovery thresholds necessitates a potentially frivolous division of what should constitute as recovery (Senn, 2001). It also entails an unwarranted normative evaluation of what sort of improvements is deemed important. In the example above, patient A's substantial gains are ignored by not achieving a level of functioning that is out potentially of reach. Researchers should keep this in mind as conclusions regarding clinically significant change can have a major impact on healthcare legislation. This particular issue is not applicable to the reliable change index which does not require the researcher to specify the point of recovery, but rather a level of change that is unlikely to occur given spontaneous fluctuation of the disease.

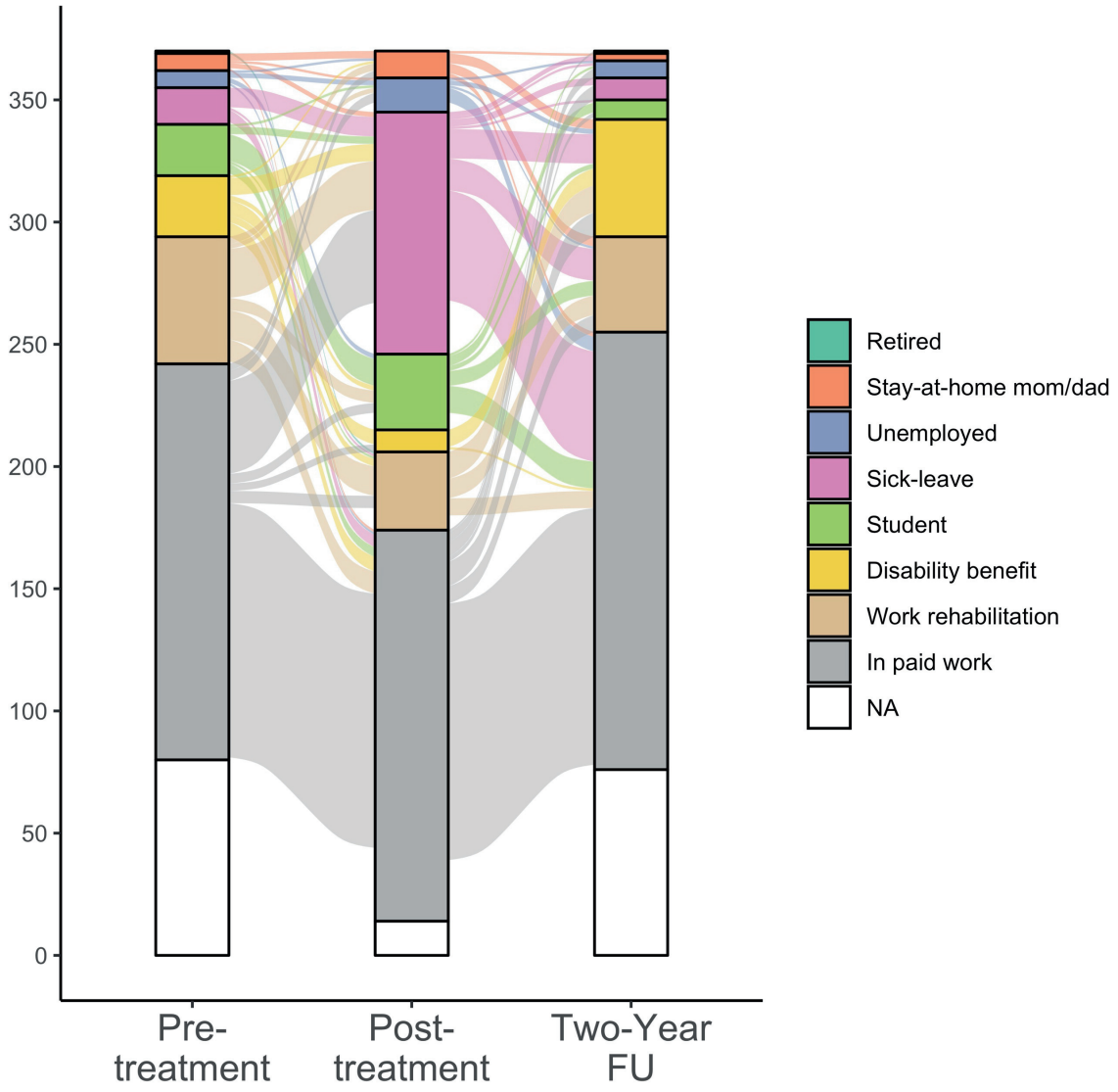
5.7.1 Changes in Occupation Functioning

The issues related to dichotomization of outcomes is conspicuous in the evaluation of occupational functioning. This phenomenon is difficult to measure on a unidimensional and continuous scale as patients often have idiosyncratic and difficult-to-define occupational status. As previously elaborated, this dichotomous presentation can mask many relevant developments. Paper I reports that occupational functioning did not significantly improve across and beyond treatment, in contrast to the large improvement seen using measures of psychopathology. This demonstrates that patients can recover from their mental illness but fail to return to work. The divergence between improvements in mental health and continued poor occupational functioning has also been demonstrated by others (Henderson et al., 2011; Løvvik et al., 2014). These results indicate that the relationship between occupational functioning and mental health is complex and intertwined with other processes. A

comprehensive approach to patient recovery necessitates a broad appreciation of the complex interplay between variables related to the patient and the work environment (Mikkelsen & Rosholm, 2018). A recent review concluded that return to work interventions that do not include workplace modifications or service coordination components are not effective for patients suffering from a mental disorder (Cullen et al., 2018). Given these findings, the unfavorable occupational findings from the NMSPOP are not surprising as this intervention was focused purely on treating mental illness.

Figure 9 illustrates the complexity of changes seen in occupational functioning. It reveals that there is no simple pattern or developmental trend that can adequately describe the data. Rather, patients seem to fluctuate between functioning and non-functioning categories. Notably, the absolute proportion of patients engaged in paid work increases slightly throughout treatment and is sustained in the follow-up period. This positive development is sourced mainly from patients in the work rehabilitation, disability, and sick-leave pretreatment column. Also note that there is a substantial amount of missing data at pretreatment and at the follow-up. The missing pretreatment data is curious as all participants received the pretreatment questionnaire. The missing data is caused by participants skipping the question completely. This reason for this might be that the question is hard to answer as the categories might not cover the patient's particular occupational situation. Notice that no generic 'other' category was provided. Given these methodological limitations and the inherent complexity in the returning-to-work phenomenon, we cannot conclude, one way or the other, whether patients show a positive or negative development following treatment.

Figure 9: Alluvial Diagram of Occupational Status



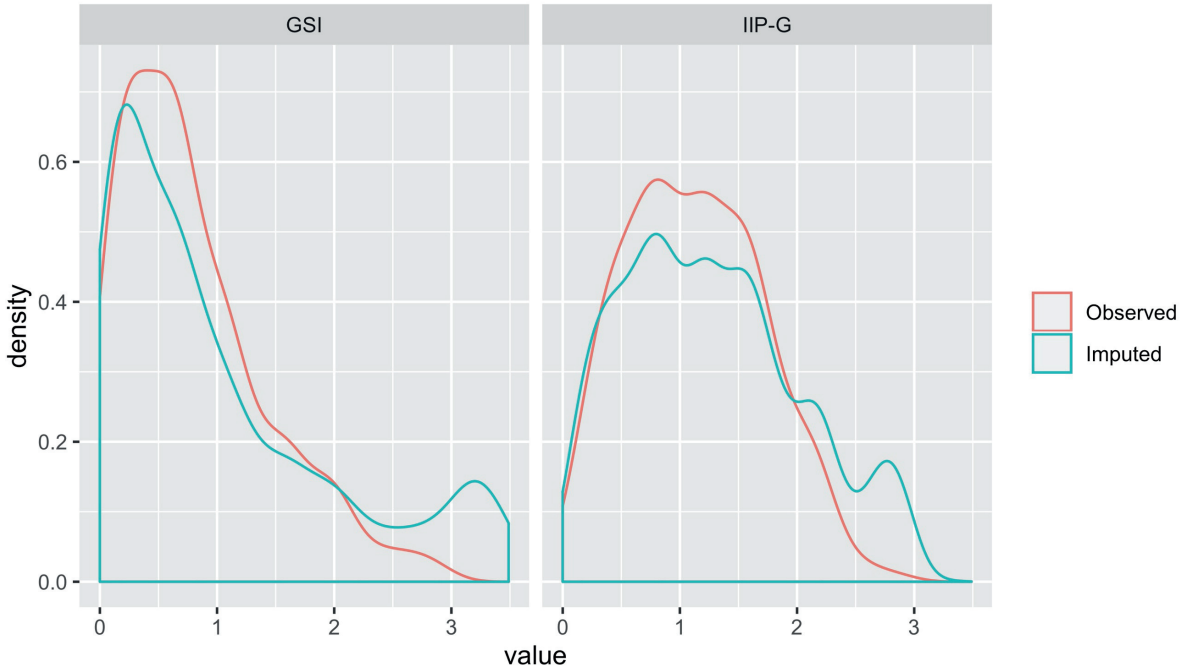
5.8 Managing Missing Data

The data presented in this thesis demonstrates a moderate amount of missing data in the follow-up portion of the trial. A total of 25 % of the sample completed no follow-up assessment. This is not uncommon in trials with long follow-up phases as patients might lose motivation or interest in participating in the study, move to a different address, change phone-number, et cetera (Graham, 2009). These are examples of benign processes from a statistical inference point-of-view. However, if the missing follow-up data is a consequence of changes in the patient's psychopathology, then it is a cause for concern. For example, if a patient

treated for depression experience a severe relapse which makes him/her unable to return questionnaires in the follow-up period. To account for this scenario, this thesis utilizes multiple imputations strategies to estimate what these missing values might have been. Specifically, Multiple Imputation by Chained Equation (MICE), which is a modern protocol for multiple imputation. MICE entails running a series of regression models where each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable is managed according to its distribution with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression. This flexibility is convenient in this thesis as many salient variables are dichotomous such as the presence of a personality disorder and occupational functioning. The MICE algorithm cycles through different imputed values for each variable, according to Rubin's (1976) rules, and converge on the best fitting parameters (Azur et al., 2011; van Ginkel et al., 2019). Missing data was imputed based on symptom severity (GSI), interpersonal problems (IIP-Global), number of positive SCID 2 Personality Disorder items, occupation functioning, and precedence of personality disorder, all measured at pretreatment. The goal of this analysis is to make an educated guess for missing GSI and IIP-Global follow-up values at the two-and-a-half-year follow-up. Ten multiply imputed datasets were utilized.

The results of this analysis are shown in figure 10. This density plot displays the imputed values with the observed follow-up values on both the GSI and IIP-Global score. The plot reveals that scores largely overlap but that there is a tendency for patients with high scores to miss their follow-up. This indicates that missing should not be considered random. Rather, patients with a predicted poor score are more likely to miss their follow-up. Note that a GSI and IIP-Global score of above 2.5 indicates severe symptoms and interpersonal problems, respectively.

Figure 10: Comparing Imputed and Observed Values for Missing Follow-Up Data



Using a complete dataset with imputed scores arguably allows us to negate this effect. We have repeated every analysis as described in paper I, II, and III and found that none of our main results and conclusions are significantly affected. A conservative conclusion is that our results are robust for a large majority of measured patients. However, we cannot exclude the possibility that causal processes, not captured in our data, have produced missing data (MNAR).

5.9 Measurement and Reliability

Any quantitative measurement of psychology faces the challenge of operationalizing and capturing a theoretical phenomenon. The phenomenon of interest is often latent in nature and cannot be directly observed. Researchers must therefore rely on approximate measures and infer the value of the true unobserved score. A measure can be said to be reliable if it is consistent while measuring a valid psychological construct. This thesis has two main continuous outcome measures, namely SCL-90-R and IIP-64. These questionnaires are aimed at capturing distinct aspects of broad psychopathological processes. The main conclusions in

this thesis collapse the total score into a single measure for both questionnaires, namely the global severity index and the global index of interpersonal problems for the SCL-90-R and IIP-64 respectively. This is a composite score that attempts to capture the overall symptoms and interpersonal problems, in contrast to individual item scores which measures specific features. The global severity index and the global index of interpersonal problems is the mean of the entire questionnaire.

The reliability of these composite measures is assessed using Cronbach's alpha, also known as coefficient alpha. This statistic captures the degree of internal consistency in the measure. It is calculated using the following formula where k is the number of items/questions and r is the mean correlation between each item: $a = \frac{kr}{(1 + (k-1)r)}$. This formula reveals that a is related to both the number of items and the overall correlation between each item. Tests with a high number of items or a high correlation between items produce a higher a. Cronbach's alpha scores can range from 0 (no consistency) to 1 (perfect consistency). A score of .7 indicates that 70 % of the variance in the measure is reliable variance. The other 30 % is due to factors that are not related to the phenomena the researcher is interested in, known as error variance. It has been suggested that a Cronbach's alpha of .7 is acceptable for exploratory research in a new field, while basic research in a developed field should aim for at least .8. It is also generally recommended that applied research, such as the work presented in this thesis, should demonstrate a Cronbach's alpha of at least .9 (Nunnally, 1975).

The main issue with modeling composite scores that have low reliability is that the models are forced to incorporate the error variance in addition to the true variance. Low reliability means that the statistical models have less precision and could overlook true effects (Type I error) or erroneously detect non-effects (Type II error) (Shrout & Rodgers, 2018). Additionally, utilizing composite scores assessed with Cronbach's alpha assumes a unidimensional construct that may not be appropriate. However, several investigations have

shown that both the GSI and IIP Global scores are highly correlated with similar broad measures of psychopathology (Bush et al., 2012; Müller et al., 2010).

The work presented in this thesis utilizes a traditional mean composite score for two reasons. Firstly, both the global severity index (SCL-90-R) and global index of interpersonal problems (IIP-64) demonstrated very high Cronbach's alpha in our sample, .97 and .93 respectively. This is an assurance that the phenomena of interest are reliably captured using the composite measures. Second, the hypothesis in this thesis are all related to research questions regarding global composite constructs, e.g. do patient improve, overall. Further research into this material should aim to use a more fine-grained approach to modeling psychiatric symptoms and interpersonal problems.

5.9 Implications and Future Directions

The results of this thesis have several clinical implications, some of which are particularly salient when treating patients that are severely afflicted with a mental illness. Clinicians should be aware that the majority of the treatment dosage literature is aimed at patients with mild, to moderate conditions, rarely encompassing patients with a personality disorder or multiple comorbid mental illnesses. This has influenced dosage recommendations as mildly afflicted patients demonstrate a more rapid treatment response. A general observation is this thesis is that clinicians should consider each patient's level of psychopathology when evaluating the appropriate treatment length. This finding is corroborated by a large empirical literature and stands in contrast to a fixed dosage perspective. Patients with mild psychopathological problems and high functioning can be expected to improve rapidly while more severe cohorts require more time.

Additionally, another key finding is that some patients do not improve with open-ended psychotherapy treatment. The data in this thesis demonstrate that patients can engage in treatment that spans several years, and still experience limited and/or unstable improvement.

These patients all report high levels of psychopathological symptoms and interpersonal problems at pretreatment, and the majority have a personality disorder. Future research should investigate treatment non-response in order to reach this segment of patients. This research can be aimed both at identifying risk factors for non-response, and also interventions to guard against non-response. Also, the observation that some patients engage in lengthy treatment without improvement is an argument for the implementation of patient monitoring and feedback tools. These tools hold promise for both increasing treatment effectiveness, as well as identifying and intervening against treatment non-response (Lambert et al., 2018).

The quest to disentangle the relationship between levels of psychopathology and treatment features is related to the development of personalized psychotherapy approaches. The development of personalized psychotherapy approaches depends on outcome research as well as clinical implementation and observation, in order to improve patient outcomes by tailoring treatment to the individual patient (Cuijpers et al., 2016; Millon & Grossman, 2007; Schneider et al., 2015). Future research should seek to move beyond coarse diagnostic categories and focus on nuanced descriptions of psychopathology and its interplay with operationally defined treatment features. This is in contrast to generic labels of psychotherapeutic orientations which therapist seldom deliver in naturalistic practice (Kazdin, 2017) and diagnostic categories with questionable validity (Fried & Nesse, 2015; McElroy et al., 2018; Selzam et al., 2018; Waszczuk et al., 2017).

Personalized approaches are related to another promising clinical innovation, namely the application of item-response theory and computerized assessment (Gibbons et al., 2016). The goal of an item-response theoretical approach is to relate each item to a corresponding latent trait. Item-response research can investigate what items are better at measuring a specific aspect of psychopathology at differing levels. For example, a score of zero vs. four on the SLC-90 item: "The idea that someone else can control your thoughts" probably has a

different sensitivity to severe psychopathology compared to the item: "A lump in your throat". A computerized monitoring routine could, in principle, gather idiosyncratic changes of mental illness were patients only answers questions that are relevant for their specific mental disorder and severity level.

6. Conclusions

There is a large body of evidence indicating that psychotherapy produces substantial and lasting positive change for individuals afflicted with a mental illness. This positive effect of psychotherapy is moderated by a host of factors relating to patient characteristics and the treatment received. The majority of outcome studies are based on an experimental treatment setting that seeks to tease out the causal relationship between the alleviation of a specific mental illness and specific treatment parameters. The treatment delivered is typically highly specified and monitored in order to achieve this goal. Less is known about the effects of psychotherapy as delivered in a naturalistic context. Naturalistic psychotherapy is differentiated from the experimental setting with respect to both the treatment delivered, allowing therapists to deliver their usual practice, and the patient cohort, which is typically more diverse and includes patients with high levels of comorbidity. The results from this thesis demonstrate that patients treated with open-ended psychotherapy, in a naturalistic context, show large positive gains in both interpersonal functioning and overall symptomatic intensity. This observation is mirrored by observer-rated diagnostic improvements which reveal that a majority of patients do not qualify for their respective SCID I and II diagnosis at posttreatment. These positive changes are stable throughout a two-and-a-half-year follow-up period. Patients with and without a personality disorder demonstrate similar improvements when defined by the overall magnitude of change. Similarly, more severely suffering patients received longer treatments, had slower rates of change but in general, received greater overall benefits. Taken as a whole, these results reveal new insights into the relationship between psychotherapy treatment conditions and outcomes. This will hopefully spark further interest in research that has the potential to improve individual patient outcomes and lives.

References

Abbass, A. A., Kisely, S. R., Town, J. M., Leichsenring, F., Driessen, E., De Maat, S., Gerber, A., Dekker, J., Rabung, S., Rusalovska, S., & Crowe, E. (2014). Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD004687.pub4

Abramowitz, J. S., Deacon, B. J., & Whiteside, S. P. H. (2019). *Exposure Therapy for Anxiety: Principles and Practice* (2nd ed.). The Guilford Press.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). American Psychiatric Publishing.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425596

American Psychological Association. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

American Psychological Association. (2015). Professional practice guidelines: Guidance for developers and users. *American Psychologist*, *70*(9), 823–831. https://doi.org/10.1037/a0039644

Andersen, P. K., & Keiding, N. (2005). Survival Analysis, Overview. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (pp. 175-198). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011815.b2a11072

Anderson, E. M., & Lambert, M. J. (2001). A survival analysis of clinically significant change in outpatient psychotherapy. *Journal of Clinical Psychology*, *57*(7), 875–888. https://doi.org/10.1002/jclp.1056

Andersson, G., Titov, N., Dear, B. F., Rozental, A., & Carlbring, P. (2019). Internet-delivered psychological treatments: from innovation to implementation. *World Psychiatry*, *18*(1), 20–28. https://doi.org/10.1002/wps.20610

Antonsen, B. T., Klungsøyr, O., Kamps, A., Hummelen, B., Johansen, M. S., Pedersen, G., Urnes, Ø., Kvarstein, E. H., Karterud, S., & Wilberg, T. (2014). Step-down versus outpatient psychotherapeutic treatment for personality disorders: 6-year follow-up of the Ullevål personality project. *BMC Psychiatry*, *14*(1), 119. https://doi.org/10.1186/1471-244X-14-119

Asay, T. P., Lambert, M. J., Gregersen, A. T., & Goates, M. K. (2002). Using patient-focused research in evaluating treatment outcome in private practice. *Journal of Clinical Psychology*, *58*(10), 1213–1225. https://doi.org/10.1002/jclp.10107

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose–effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, *77*(2), 203–211. https://doi.org/10.1037/a0015235

Barkham, M., Culverwell, A., Spindler, K., & Twigg, E. (2005). The CORE-OM in an older adult population: Psychometric status, acceptability, and feasibility. *Aging & Mental Health*, *9*(3), 235–245. https://doi.org/10.1080/13607860500090052

Barkham, Michael, Connell, J., Stiles, W. B., Miles, J. N. V., Margison, F., Evans, C., & Mellor-Clark, J. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology*, *74*(1), 160–167. https://doi.org/10.1037/0022-006X.74.1.160

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. *Journal of Consulting and Clinical Psychology*, *49*(2), 147–155. https://doi.org/10.1037/0022-006X.49.2.147

Bateman, A., & Fonagy, P. (2008). 8-Year Follow-Up of Patients Treated for Borderline Personality Disorder: Mentalization-Based Treatment Versus Treatment as Usual. *American Journal of Psychiatry*, *165*(5), 631–638. https://doi.org/10.1176/appi.ajp.2007.07040636

Bateman, A. W., Gunderson, J., & Mulder, R. (2015). Treatment of personality disorder. *The Lancet*, *385*(9969), 735–743. https://doi.org/10.1016/S0140-6736(14)61394-5

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

Beckwith, H., Moran, P. F., & Reilly, J. (2014). Personality disorder prevalence in psychiatric outpatients: A systematic literature review: Personality disorder prevalence in psychiatric outpatients. *Personality and Mental Health*, *8*(2), 91–101. https://doi.org/10.1002/pmh.1252

Berg, H., & Slaattelid, R. (2017). Facts and values in psychotherapy-A critique of the empirical reduction of psychotherapy within evidence-based practice. *Journal of Evaluation in Clinical Practice*, *23*(5), 1075–1080. https://doi.org/10.1111/jep.12739

Bernhard, J. M., & Goodyear, R. K. (2019). *Fundamentals of Clinical Supervision*. The Merill Counseling Series.

Beutler, L. E., Castonguay, L. G., & Follette, W. C. (2006). Therapeutic factors in dysphoric disorders. *Journal of Clinical Psychology*, *62*(6), 639–647. https://doi.org/10.1002/jclp.20260

Bjerke, E., Hansen, R. S., Solbakken, O. A., & Monsen, J. T. (2011). Interpersonal problems among 988 Norwegian psychiatric outpatients. A study of pretreatment self-reports. *Comprehensive Psychiatry*, *52*(3), 273–279. https://doi.org/10.1016/j.comppsych.2010.07.004

Bohart, A. C., & Greaves, W. (2013). Client Contributions to Therapy Process and Outcomes. In *Bergin & Garfields Handbook of Psychotherapy and Behavior Change* (pp. 219–257). John Wiley & Sons.

Breuer, J., & Freud, S. (2010). *Studies on Hysteria*. Martino Fine Books. (Original work published 1895)

Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J. (2001). Pushing the Quality Envelope: A New Outcomes Management System. *Psychiatric Services*, *52*(7), 925–934. https://doi.org/10.1176/appi.ps.52.7.925

Brown, J. (2015). Specific Techniques Vs. Common Factors? Psychotherapy Integration and its Role in Ethical Practice. *American Journal of Psychotherapy*, *69*(3), 301–316. https://doi.org/10.1176/appi.psychotherapy.2015.69.3.301

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147–158. https://doi.org/10.1037/0033-2909.101.1.147

Budd, R., & Hughes, I. (2009). The Dodo Bird Verdict-controversial, inevitable and important: A commentary on 30 years of meta-analyses: The Dodo Bird Verdict: A Commentary. *Clinical Psychology & Psychotherapy*, *16*(6), 510–522. https://doi.org/10.1002/cpp.648

Burstein, L. (1980). The Analysis of Multilevel Data in Educational Research and Evaluation. *Review of Research in Education*, *8*, 158-233. https://doi.org/10.2307/1167125

Bush, A. L., Patel, A. B., Allen, J. G., Teal, C., Latini, D. M., Ellis, T. E., Herrera, S., & Frueh, B. C. (2012). Factor Structure and Convergent Validity of the Inventory of Interpersonal Problems in an Inpatient Setting: *Journal of Psychiatric Practice*, *18*(3), 145–158. https://doi.org/10.1097/01.pra.0000415072.36121.2d

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation*, *1986*(31), 67–77. https://doi.org/10.1002/ev.1434

Campbell, L. F., Norcross, J. C., Vasquez, M. J. T., & Kaslow, N. J. (2013). Recognition of psychotherapy effectiveness: The APA resolution. *Psychotherapy*, *50*(1), 98–101. https://doi.org/10.1037/a0031817

Chambers, C. (2017). *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton University Press.

Chambless, D. L. (1995). Training in and Dissemination of Empirically-Validated Psychological Treatments: Report and Recommendations. *The Clinical Psychologist, 48*(1). 3-23. https://doi.org/10.1037/e554972011-003

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., & Muller, K. E. (2010). Real longitudinal data analysis for real people: Building a good enough mixed model. *Statistics in Medicine*, *29*(4), 504–520. https://doi.org/10.1002/sim.3775

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry*, *23*(4), 318–327. https://doi.org/10.3109/09540261.2011.606803

Clarkin, J. F., Levy, K. N., Lenzenweger, M. F., & Kernberg, O. F. (2007). Evaluating Three Treatments for Borderline Personality Disorder: A Multiwave Study. *American Journal of Psychiatry*, *164*(6), 922–928. https://doi.org/10.1176/ajp.2007.164.6.922

Cristea, I. A., Gentili, C., Cotet, C. D., Palomba, D., Barbui, C., & Cuijpers, P. (2017). Efficacy of Psychotherapies for Borderline Personality Disorder: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *74*(4), 319–328. https://doi.org/10.1001/jamapsychiatry.2016.4287

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., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2019). Is psychotherapy effective? Pretending everything is fine will not help the field forward. *Epidemiology and Psychiatric Sciences*, *28*(03), 356–357. https://doi.org/10.1017/S204579601800080X

Cuijpers, P., Reijnders, M., & Huibers, M. J. H. (2019). The Role of Common Factors in Psychotherapy Outcomes. *Annual Review of Clinical Psychology*, *15*(1), 207–231. https://doi.org/10.1146/annurev-clinpsy-050718-095424

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *The British Journal of Psychiatry*, *196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

Cuijpers, P., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., & Coyne, J. (2014). Comparison of psychotherapies for adult depression to pill placebo control groups: A meta-analysis. *Psychological Medicine*, *44*(4), 685–695. https://doi.org/10.1017/S0033291713000457

Cullen, K. L., Irvin, E., Collie, A., Clay, F., Gensby, U., Jennings, P. A., Hogg-Johnson, S., Kristman, V., Laberge, M., McKenzie, D., Newnam, S., Palagyi, A., Ruseckaite, R., Sheppard, D. M., Shourie, S., Steenstra, I., Van Eerd, D., & Amick, B. C. (2018). Effectiveness of Workplace Interventions in Return-to-Work for Musculoskeletal,

Pain-Related and Mental Health Conditions: An Update of the Evidence and Messages for Practitioners. *Journal of Occupational Rehabilitation*, *28*(1), 1–15. https://doi.org/10.1007/s10926-016-9690-x

De Geest, R. M., & Meganck, R. (2019). How Do Time Limits Affect Our Psychotherapies? A Literature Review. *Psychologica Belgica*, *59*(1), 206–226. https://doi.org/10.5334/pb.475

de Mello, M. F., de Jesus Mari, J., Bacaltchuk, J., Verdeli, H., & Neugebauer, R. (2005). A systematic review of research findings on the efficacy of interpersonal therapy for depressive disorders. *European Archives of Psychiatry and Clinical Neuroscience*, *255*(2), 75–82. https://doi.org/10.1007/s00406-004-0542-x

Derogatis, L. R., & Unger, R. (2010). Symptom Checklist-90-Revised. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology*. John Wiley & Sons. https://doi.org/10.1002/9780470479216.corpsy0970

Dragioti, E., Karathanos, V., Gerdle, B., & Evangelou, E. (2017). Does psychotherapy work? An umbrella review of meta-analyses of randomized controlled trials. *Acta Psychiatrica Scandinavica*, *136*(3), 236–246. https://doi.org/10.1111/acps.12713

Draper, M. R., Jennings, J., Baron, A., Erdur, O., & Shankar, L. (2002). Time-Limited Counseling Outcome in a Nationwide College Counseling Center Sample. *Journal of College Counseling*, *5*(1), 26–38. https://doi.org/10.1002/j.2161-1882.2002.tb00204.x

Dunn, C. (2016). *Ethical Issues in Mental Illness*. Routledge. https://doi.org/10.4324/9781315256115

Eisenbarth, H., Godinez, D., du Pont, A., Corley, R. P., Stallings, M. C., & Rhee, S. H. (2019). The influence of stressful life events, psychopathy, and their interaction on internalizing and externalizing psychopathology. *Psychiatry Research*, *272*, 438–446. https://doi.org/10.1016/j.psychres.2018.12.145

Elliott, R., Watson, J. C., Goldman, R. N., & Greenberg, L. S. (2004). Empty chair work for unfinished interpersonal issues. In R. Elliott, J. C. Watson, R. N. Goldman, & L. S. Greenberg, *Learning Emotion-Focused Therapy: The Process-Experiential Approach to Change.* (pp. 243–265). American Psychological Association. https://doi.org/10.1037/10725-012

Ellsworth, J. R., Lambert, M. J., & Johnson, J. (2006). A comparison of the Outcome Questionnaire-45 and Outcome Questionnaire-30 in classification and prediction of treatment outcome. *Clinical Psychology & Psychotherapy*, *13*(6), 380–391. https://doi.org/10.1002/cpp.503

Enders, C. K. (2011). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*(4), 267–288. https://doi.org/10.1037/a0025579

Eysenck, H. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*(5), 319–324. https://doi.org/10.1037/h0063633

Eysenck, H. (1966). *The Effects of Psychotherapy*. International Science Press.

Falkenström, F., Josefsson, A., Berggren, T., & Holmqvist, R. (2016). How much therapy is enough? Comparing dose-effect and good-enough models in two different settings. *Psychotherapy*, *53*, 130–139. https://doi.org/10.1037/pst0000039

Firth, J., Torous, J., Carney, R., Newby, J., Cosco, T. D., Christensen, H., & Sarris, J. (2018). Digital Technologies in the Treatment of Anxiety: Recent Innovations and Future Directions. *Current Psychiatry Reports*, *20*(6), 44. https://doi.org/10.1007/s11920-018-0910-2

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis*. John Wiley & Sons. https://doi.org/10.1002/9781119513469

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*(4), 316–340. https://doi.org/10.1037/pst0000172

Frank, J., David. (1973). *Persuasion and Healing: A Comparative Study of Psychotherapy*. John Hopkins University Press.

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M. H., Moreau, Y., Murphy, S. A., Przytycka, T. M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Stekhoven, D., Sun, J., Weber, A., Ziemek, D., & Zupan, B. (2018). From hype to reality: Data science enabling personalized medicine. *BMC Medicine*, *16*(1), 150. https://doi.org/10.1186/s12916-018-1122-7

Furnham, A., Milner, R., Akhtar, R., & Fruyt, F. D. (2014). A Review of the Measures Designed to Assess DSM-5 Personality Disorders. *Psychology*, *5*(14), 1646–1686. https://doi.org/10.4236/psych.2014.514175

Gabbard, G. O. (2017). *Long-term Psychodynamic Psychotherapy: A Basic Text* (3rd ed.). American Psychiatric Publishing.

Gelman, A. (2005). Analysis of Variance: Why It Is More Important Than Ever. *The Annals of Statistics, 33*(1), 1–53. https://doi.org/10.1214/009053604000001048

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized Adaptive Diagnosis and Testing of Mental Health Disorders. *Annual Review of Clinical Psychology*, *12*(1), 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gkeredakis, E., Swan, J., Powell, J., Nicolini, D., Scarbrough, H., Roginski, C., Taylor-Phillips, S., & Clarke, A. (2011). Mind the gap: Understanding utilization of evidence

and policy in health care management practice. *Journal of Health Organization and Management*, *25*(3), 298–314. https://doi.org/10.1108/14777261111143545

Gonzalez, D. M. (2016). Client variables and psychotherapy outcomes. In D. J. Cain, K. Keenan, & S. Rubin (Eds.), *Humanistic psychotherapies: Handbook of research and practice* (2nd ed.). (pp. 455–482). American Psychological Association. https://doi.org/10.1037/14775-015

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Green, J. L., Lowry, J. L., & Kopta, S. M. (2003). College Students versus College Counseling Center Clients: What Are the Differences? *Journal of College Student Psychotherapy*, *17*(4), 25–37. https://doi.org/10.1300/J035v17n04_05

Gurka, M. J. (2006). Selecting the Best Linear Mixed Model Under REML. *The American Statistician*, *60*(1), 19–26. https://doi.org/10.1198/000313006X90396

Hamer, R. M., & Simpson, P. M. (2009). Last Observation Carried Forward Versus Mixed Models in the Analysis of Psychiatric Clinical Trials. *American Journal of Psychiatry*, *166*, 639–641. https://doi.org/10.1176/appi.ajp.2009.09040458

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2006). The Psychotherapy Dose-Response Effect and Its Implications for Treatment Delivery Services. *Clinical Psychology: Science and Practice*, *9*, 329–343. https://doi.org/10.1093/clipsy.9.3.329

Henderson, M., Harvey, S., Øverland, S., Mykletun, A., & Hotopf, M. (2011). Work and common psychiatric disorders. *Journal of the Royal Society of Medicine*, *104*(5), 198–207. https://doi.org/10.1258/jrsm.2011.100231

Henry, P. J. (2008). Student Sampling as a Theoretical Problem. *Psychological Inquiry*, *19*(2), 114–126. https://doi.org/10.1080/10478400802049951

Hoffman, I. Z. (2009). Doublethinking Our Way to “Scientific” Legitimacy: The Desiccation of Human Experience. *Journal of the American Psychoanalytic Association*, *57*(5), 1043–1069. https://doi.org/10.1177/0003065109343925

Horowitz, L. M., Alden, L. E., Wiggins, J. S., & Pincus, A. L. (1979). *Inventory of interpersonal problems*. American Psychological Association.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist*, *41*(2), 159–164. https://doi.org/10.1037/0003-066X.41.2.159

Imel, Z. E., Caperton, D. D., Tanana, M., & Atkins, D. C. (2017). Technology-enhanced human interaction in psychotherapy. *Journal of Counseling Psychology*, *64*(4), 385–393. https://doi.org/10.1037/cou0000213

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(4), 1195–1199. https://doi.org/10.1037/a0015665

Jørgensen, C. R. (2019). Medical Versus Dynamic-Relational Model of Psychotherapy. In C. R. Jørgensen, *The Psychotherapeutic Stance* (pp. 29–39). Springer International Publishing. https://doi.org/10.1007/978-3-030-20437-2_3

Jørgensen, C. R., Bøye, R., Andersen, D., Døssing Blaabjerg, A. H., Freund, C., Jordet, H., & Kjølbye, M. (2014). Eighteen months post-treatment naturalistic follow-up study of mentalization-based therapy and supportive group treatment of borderline personality disorder: Clinical outcomes and functioning. *Nordic Psychology*, *66*(4), 254–273. https://doi.org/10.1080/19012276.2014.963649

Kanter, J. W., & Puspitasari, A. J. (2016). Global dissemination and implementation of behavioural activation. *The Lancet*, *388*(10047), 843–844. https://doi.org/10.1016/S0140-6736(16)31131-X

Karlin, B. E., & Cross, G. (2014). From the laboratory to the therapy room: National dissemination and implementation of evidence-based psychotherapies in the U.S. Department of Veterans Affairs Health Care System. *American Psychologist*, *69*(1), 19–33. https://doi.org/10.1037/a0033888

Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy*, *88*, 7–18. https://doi.org/10.1016/j.brat.2016.06.004

Kirsch, I., Wampold, B., & Kelley, J. M. (2016). Controlling for the placebo effect in psychotherapy: Noble quest or tilting at windmills? *Psychology of Consciousness: Theory, Research, and Practice*, *3*(2), 121–131. https://doi.org/10.1037/cns0000065

Knapstad, M., Nordgreen, T., & Smith, O. R. F. (2018). Prompt mental health care, the Norwegian version of IAPT: Clinical outcomes and predictors of change in a multicenter cohort study. *BMC Psychiatry*, *18*(1), 260. https://doi.org/10.1186/s12888-018-1838-0

Knekt, P., Virtala, E., Härkänen, T., Vaarama, M., Lehtonen, J., & Lindfors, O. (2016). The outcome of short- and long-term psychotherapy 10 years after start of treatment. *Psychological Medicine*, *46*(6), 1175–1188. https://doi.org/10.1017/S0033291715002718

Kolden, G. G. (1991). The generic model of psychotherapy: An empirical investigation of patterns of process and outcome relationships. *Psychotherapy Research*, *1*(1), 62–73. https://doi.org/10.1080/10503309112331334071

Kopta, S. M. (2003). The dose—effect relationship in psychotherapy: A defining achievement for Dr. Kenneth Howard. *Journal of Clinical Psychology*, *59*, 727–733. https://doi.org/10.1002/jclp.10167

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin*, *143*(2), 142–186. https://doi.org/10.1037/bul0000069

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lambert, M. J. (2017). Maximizing Psychotherapy Outcome beyond Evidence-Based Medicine. *Psychotherapy and Psychosomatics*, *86*(2), 80–89. https://doi.org/10.1159/000455170

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy: An International Journal of Theory and Practice*, *3*(4), 249–258.

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy*, *55*(4), 520–537. https://doi.org/10.1037/pst0000167

Langeland, E., Riise, T., Hanestad, B. R., Nortvedt, M. W., Kristoffersen, K., & Wahl, A. K. (2006). The effect of salutogenic treatment principles on coping with mental health problems. *Patient Education and Counseling*, *62*(2), 212–219. https://doi.org/10.1016/j.pec.2005.07.004

Laurenceau, J.-P., Hayes, A. M., & Feldman, G. C. (2007). Some methodological and statistical issues in the study of change processes in psychotherapy. *Clinical Psychology Review*, *27*(6), 682–695. https://doi.org/10.1016/j.cpr.2007.01.007

Leichsenring, F., & Rabung, S. (2011). Long-term psychodynamic psychotherapy in complex mental disorders: Update of a meta-analysis. *British Journal of Psychiatry*, *199*(1), 15–22. https://doi.org/10.1192/bjp.bp.110.082776

Leuzinger-Bohleber, M., Hautzinger, M., Fiedler, G., Keller, W., Bahrke, U., Kallenbach, L., Kaufhold, J., Ernst, M., Negele, A., Schoett, M., Küchenhoff, H., Günther, F., Rüger, B., & Beutel, M. (2019). Outcome of Psychoanalytic and Cognitive-Behavioural Long-Term Therapy with Chronically Depressed Patients: A Controlled Trial with Preferential and Randomized Allocation. *The Canadian Journal of Psychiatry*, *64*(1), 47–58. https://doi.org/10.1177/0706743718780340

Lilienfeld, S. O., McKay, D., & Hollon, S. D. (2018). Why randomised controlled trials of psychological treatments are still essential. *The Lancet Psychiatry*, *5*(7), 536–538. https://doi.org/10.1016/S2215-0366(18)30045-2

Lincoln, T. M., Jung, E., Wiesjahn, M., & Schlier, B. (2016). What is the minimal dose of cognitive behavior therapy for psychosis? An approximation using repeated assessments over 45 sessions. *European Psychiatry*, *38*, 31–39. https://doi.org/10.1016/j.eurpsy.2016.05.004

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181–1209. https://doi.org/10.1037/0003-066X.48.12.1181

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Livesley, J., W., & Larstone, R. (2018). Empirically based treatments. In J.W. Livesley & R. Larstone (Eds.) *Handbook of Personality Disorders* (2nd ed., pp. 481–487). The Guilford Press.

Lobbestael, J., Leurgans, M., & Arntz, A. (2011). Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clinical Psychology & Psychotherapy*, *18*(1), 75–79. https://doi.org/10.1002/cpp.693

Lockard, A. J., Hayes, J. A., McAleavey, A. A., & Locke, B. D. (2012). Change in Academic Distress: Examining Differences Between a Clinical and Nonclinical Sample of College Students. *Journal of College Counseling*, *15*(3), 233–246. https://doi.org/10.1002/j.2161-1882.2012.00018.x

Lorentzen, S., & Høglend, P. (2004). Predictors of Change during Long-Term Analytic Group Psychotherapy. *Psychotherapy and Psychosomatics*, *73*(1), 25–35. https://doi.org/10.1159/000074437

Løvvik, C., Shaw, W., Øverland, S., & Reme, S. E. (2014). Expectations and illness perceptions as predictors of benefit recipiency among workers with common mental disorders: Secondary analysis from a randomised controlled trial. *BMJ Open*, *4*(3), e004321. https://doi.org/10.1136/bmjopen-2013-004321

Maat, S. de, Dekker, J., Schoevers, R., & Jonghe, F. de. (2007). The effectiveness of long-term psychotherapy: Methodological research issues. *Psychotherapy Research*, *17*(1), 59–65. https://doi.org/10.1080/10503300600607605

Magnusson, K., Andersson, G., & Carlbring, P. (2018). The consequences of ignoring therapist effects in trials with longitudinal data: A simulation study. *Journal of Consulting and Clinical Psychology*, *86*(9), 711–725. https://doi.org/10.1037/ccp0000333

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McAleavey, A. A., Youn, S. J., Xiao, H., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2019). Effectiveness of routine psychotherapy: Method matters. *Psychotherapy Research*, *29*(2), 139–156. https://doi.org/10.1080/10503307.2017.1395921

McElroy, E., Fearon, P., Belsky, J., Fonagy, P., & Patalay, P. (2018). Networks of Depression and Anxiety Symptoms Across Development. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*(12), 964–973. https://doi.org/10.1016/j.jaac.2018.05.027

McMain, S. F., Links, P. S., Gnam, W. H., Guimond, T., Cardish, R. J., Korman, L., & Streiner, D. L. (2009). A Randomized Trial of Dialectical Behavior Therapy Versus General Psychiatric Management for Borderline Personality Disorder. *American Journal of Psychiatry*, *166*(12), 1365–1374. https://doi.org/10.1176/appi.ajp.2009.09010039

Meares, R., Stevenson, J., & Comerford, A. (1999). Psychotherapy with Borderline Patients: I. A Comparison Between Treated and Untreated Cohorts. *Australian & New Zealand Journal of Psychiatry*, *33*(4), 467–472. https://doi.org/10.1080/j.1440-1614.1999.00594.x

Mikkelsen, M. B., & Rosholm, M. (2018). Systematic review and meta-analysis of interventions aimed at enhancing return to work for sick-listed workers with common mental disorders, stress-related disorders, somatoform disorders and personality disorders. *Occupational and Environmental Medicine*, *75*(9), 675–686. https://doi.org/10.1136/oemed-2018-105073

Millon, T., & Grossman, S. (2007). *Moderating Severe Personality Disorders: A Personalized Psychotherapy Approach*. John Wiley & Sons. https://doi.org/10.1002/9781118269893

Milton, M. (2002). Evidence-Based Practice: Issues for Psychotherapy. *Psychoanalytic Psychotherapy*, *16*(2), 160–172. https://doi.org/10.1080/14749730210133429

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology*, *76*(1), 116–124. https://doi.org/10.1037/0022-006X.76.1.116

Müller, J. M., Postert, C., Beyer, T., Furniss, T., & Achtergarde, S. (2010). Comparison of Eleven Short Versions of the Symptom Checklist 90-Revised (SCL-90-R) for Use in the Assessment of General Psychopathology. *Journal of Psychopathology and Behavioral Assessment*, *32*(2), 246–254. https://doi.org/10.1007/s10862-009-9141-5

Munder, T., Brütsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review*, *33*(4), 501–511. https://doi.org/10.1016/j.cpr.2013.02.002

Nathan, P. E., Stuart, S. P., & Dolan, S. L. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *Psychological Bulletin*, *126*(6), 964–981. https://doi.org/10.1037/0033-2909.126.6.964

Newby, J. M., McKinnon, A., Kuyken, W., Gilbody, S., & Dalgleish, T. (2015). Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. *Clinical Psychology Review*, *40*, 91–110. https://doi.org/10.1016/j.cpr.2015.06.002

Newman, M. G., & Stiles, W. B. (2006). Therapeutic factors in treating anxiety disorders. *Journal of Clinical Psychology*, *62*(6), 649–659. https://doi.org/10.1002/jclp.20262

Nielsen, E., Kirtley, O. J., & Townsend, E. (2017). "Great powers and great responsibilities": A brief comment on "A brief mobile app reduces nonsuicidal and suicidal self-injury: Evidence from three randomized controlled trials" (Franklin et al., 2016). *Journal of Consulting and Clinical Psychology*, *85*(8), 826–830. https://doi.org/10.1037/ccp0000189

Nissen, T., & Wynn, R. (2014). The history of the case report: A selective review. *Journal of the Royal Society Medicine Open*, *5*(4). https://doi.org/10.1177/2054270414523410

Nissen-Lie, H. A., Monsen, J. T., Ulleberg, P., & Rønnestad, M. H. (2013). Psychotherapists' self-reports of their interpersonal functioning and difficulties in practice as predictors of patient outcome. *Psychotherapy Research*, *23*(1), 86–104. https://doi.org/10.1080/10503307.2012.735775

Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work II. *Psychotherapy*, *48*(1), 4–8. https://doi.org/10.1037/a0022180

Nunnally, J. C. (1975). Psychometric theory 25 years ago and now. *Educational Researcher*, *4*(10), 14–20. https://doi.org/10.2307/1175619

Oud, M., Arntz, A., Hermens, M. L., Verhoef, R., & Kendall, T. (2018). Specialized psychotherapies for adults with borderline personality disorder: A systematic review and meta-analysis. *Australian & New Zealand Journal of Psychiatry*, *52*(10), 949–961. https://doi.org/10.1177/0004867418791257

Owen, J. J., Adelson, J., Budge, S., Kopta, S. M., & Reese, R. J. (2016). Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research*, *26*, 22–30. https://doi.org/10.1080/10503307.2014.966346

Parry, G., Cape, J., & Pilling, S. (2003). Clinical practice guidelines in clinical psychology and psychotherapy. *Clinical Psychology & Psychotherapy*, *10*(6), 337–351. https://doi.org/10.1002/cpp.381

Patterson, B., Boyle, M. H., Kivlenieks, M., & Van Ameringen, M. (2016). The use of waitlists as control conditions in anxiety disorders research. *Journal of Psychiatric Research*, *83*, 112–120. https://doi.org/10.1016/j.jpsychires.2016.08.015

Peeters, F., Huibers, M., Roelofs, J., van Breukelen, G., Hollon, S. D., Markowitz, J. C., van Os, J., & Arntz, A. (2013). The clinical effectiveness of evidence-based interventions for depression: A pragmatic trial in routine practice. *Journal of Affective Disorders*, *145*(3), 349–355. https://doi.org/10.1016/j.jad.2012.08.022

Perry, J. C., & Bond, M. (2009). The Sequence of Recovery in Long-Term Dynamic Psychotherapy: *The Journal of Nervous and Mental Disease*, *197*(12), 930–937. https://doi.org/10.1097/NMD.0b013e3181c29a0f

Philips, B. (2009). Comparing apples and oranges: How do patient characteristics and treatment goals vary between different forms of psychotherapy? *Psychology and Psychotherapy: Theory, Research and Practice*, *82*(3), 323–336. https://doi.org/10.1348/147608309X431491

Puolakanaho, A., Tolvanen, A., Kinnunen, S. M., & Lappalainen, R. (2020). A psychological flexibility -based intervention for Burnout: A randomized controlled trial. *Journal of Contextual Behavioral Science*, *15*, 52–67. https://doi.org/10.1016/j.jcbs.2019.11.007

Reese, R. J., Toland, M. D., & Hopkins, N. B. (2011). Replicating and extending the good-enough level model of change: Considering session frequency. *Psychotherapy Research*, *21*(5), 608–619. https://doi.org/10.1080/10503307.2011.598580

Rieken, B., & Gelo, O. C. G. (2015). The Philosophy of Psychotherapy Science: Mainstream and Alternative Views. In O. C. G. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy Research* (pp. 67–92). Springer Vienna. https://doi.org/10.1007/978-3-7091-1382-0_4

Robinson, L., Delgadillo, J., & Kellett, S. (2019). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research 30*(1), 1–18. https://doi.org/10.1080/10503307.2019.1566676

Rogers, C. (1951). *Client-Centered Therapy: Its Current Practice, Implications and Theory*. Constable & Robinson Ltd.

Rosenthal, D., & Frank, J. D. (1956). Psychotherapy and the placebo effect. *Psychological Bulletin*, *53*(4), 294–302. https://doi.org/10.1037/h0044068

Rosenzweig, S. (1954). A transvaluation of psychotherapy: A reply to Hans Eysenck. *The Journal of Abnormal and Social Psychology*, *49*(2), 298–304. https://doi.org/10.1037/h0061172

Rozental, A., Castonguay, L., Dimidjian, S., Lambert, M., Shafran, R., Andersson, G., & Carlbring, P. (2018). Negative effects in psychotherapy: Commentary and recommendations for future research and clinical practice. *BJPsych Open*, *4*(4), 307–312. https://doi.org/10.1192/bjo.2018.42

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schneider, R. L., Arch, J. J., & Wolitzky-Taylor, K. B. (2015). The state of personalized treatment for anxiety disorders: A systematic review of treatment moderators. *Clinical Psychology Review*, *38*, 39–54. https://doi.org/10.1016/j.cpr.2015.02.004

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. Wiley-Interscience.

Selzam, S., Coleman, J. R. I., Caspi, A., Moffitt, T. E., & Plomin, R. (2018). A polygenic p factor for major psychiatric disorders. *Translational Psychiatry*, *8*(1), 205. https://doi.org/10.1038/s41398-018-0217-4

Senn, S. (2001). Individual Therapy: New Dawn or False Dawn? *Drug Information Journal*, *35*, 1479–1494. https://doi.org/10.1177/009286150103500443

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D., Jorm, A. F., Lyons, L. C., Nietzel, M. T., Robinson, L., Prout, H. T., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works in clincally representative conditions. *Journal of Consulting and Clinical Psychology*, *65*(3), 355–365. https://doi.org/10.1037/0022-006X.65.3.355

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*(4), 512–529. https://doi.org/10.1037/0033-2909.126.4.512

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, *92*(3), 581–604. https://doi.org/10.1037/0033-2909.92.3.581

Sharp, C., & Kalpakci, A. (2015). Mentalization in borderline personality disorder: From bench to bedside. *Personality Disorders: Theory, Research, and Treatment*, *6*(4), 347–355. https://doi.org/10.1037/per0000106

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American Psychologist*, *65*(2), 98–109. https://doi.org/10.1037/a0018378

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications.

Staines, G. L., & Cleland, C. M. (2007). Bias in Meta-Analytic Estimates of the Absolute Efficacy of Psychotherapy. *Review of General Psychology*, *11*(4), 329–347. https://doi.org/10.1037/1089-2680.11.4.329

Steele, F. (2008). Multilevel models for longitudinal data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(1), 5–19. https://doi.org/10.1111/j.1467-985X.2007.00509.x

Steinert, C., Munder, T., Rabung, S., Hoyer, J., & Leichsenring, F. (2017). Psychodynamic Therapy: As Efficacious as Other Empirically Supported Treatments? A Meta-Analysis Testing Equivalence of Outcomes. *American Journal of Psychiatry*, *174*(10), 943–953. https://doi.org/10.1176/appi.ajp.2017.17010057

Steinert, C., Stadter, K., Stark, R., & Leichsenring, F. (2017). The Effects of Waiting for Treatment: A Meta-Analysis of Waitlist Control Groups in Randomized Controlled Trials for Social Anxiety Disorder: The effects of waiting for treatment. *Clinical Psychology & Psychotherapy*, *24*(3), 649–660. https://doi.org/10.1002/cpp.2032

Stewart, R. E., & Chambless, D. L. (2009). Cognitive–behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology*, *77*(4), 595–606. https://doi.org/10.1037/a0016032

Stiles, W. B., & Horvath, A. O. (2017). Appropriate responsiveness as a contribution to therapist effects. In L. G. Castonguay & C. E. Hill (Eds.), *How and why are some*

therapists better than others? Understanding therapist effects. (pp. 71–84). American Psychological Association. https://doi.org/10.1037/0000034-005

Stiles, W. B., & Shapiro, D. A. (1989). Abuse of the drug metaphor in psychotherapy process-outcome research. *Clinical Psychology Review*, *9*(4), 521–543. https://doi.org/10.1016/0272-7358(89)90007-X

Strupp, H. H. (1963). The outcome problem in psychotherapy revisited. *Psychotherapy: Theory, Research & Practice*, *1*(1), 1–13. https://doi.org/10.1037/h0088565

Tasca, G. A., & Gallop, R. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. *Psychotherapy Research*, *19*(4–5), 429–437. https://doi.org/10.1080/10503300802641444

Taylor, D. J., & Pruiksma, K. E. (2014). Cognitive and behavioural therapy for insomnia (CBT-I) in psychiatric populations: A systematic review. *International Review of Psychiatry*, *26*(2), 205–213. https://doi.org/10.3109/09540261.2014.902808

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically Supported Treatment: Recommendations for a New Model. *Clinical Psychology: Science and Practice*, *22*(4), 317–338. https://doi.org/10.1111/cpsp.12122

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The Prevalence of Personality Disorders in a Community Sample. *Archives of General Psychiatry*, *58*(6), 590. https://doi.org/10.1001/archpsyc.58.6.590

Tracey, T. J. G., Wampold, B. E., Lichtenberg, J. W., & Goodyear, R. K. (2014). Expertise in psychotherapy: An elusive goal? *American Psychologist*, *69*(3), 218–229. https://doi.org/10.1037/a0035099

van Beljouw, I. M., Verhaak, P. F., Cuijpers, P., van Marwijk, H. W., & Penninx, B. W. (2010). The course of untreated anxiety and depression, and determinants of poor one-

year outcome: A one-year cohort study. *BMC Psychiatry*, *10*(1), 86. https://doi.org/10.1186/1471-244X-10-86

van Gelderen, M. J., Nijdam, M. J., & Vermetten, E. (2018). An Innovative Framework for Delivering Psychotherapy to Patients With Treatment-Resistant Posttraumatic Stress Disorder: Rationale for Interactive Motion-Assisted Therapy. *Frontiers in Psychiatry*, *9*, 176-182. https://doi.org/10.3389/fpsyt.2018.00176

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2019). Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *Journal of Personality Assessment*, 1–12. https://doi.org/10.1080/00223891.2018.1530680

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology*, *66*(2), 231–239. https://doi.org/10.1037/0022-006X.66.2.231

Wallerstein, R. S. (2001). The generations of psychotherapy research: An overview. *Psychoanalytic Psychology*, *18*(2), 243–267. https://doi.org/10.1037/0736-9735.18.2.243

Wampold, B. E. (2015). *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work* (2nd ed.). Routledge. https://doi.org/10.4324/9780203582015

Wampold, B. E. (2019). *The basics of psychotherapy: An introduction to theory and practice* (2nd ed.). American Psychological Association. https://doi.org/10.1037/0000117-000

Waszczuk, M. A., Zimmerman, M., Ruggero, C., Li, K., MacNamara, A., Weinberg, A., Hajcak, G., Watson, D., & Kotov, R. (2017). What do clinicians treat: Diagnoses or symptoms? The incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns.

Comprehensive Psychiatry, *79*, 80–88. https://doi.org/10.1016/j.comppsych.2017.04.004

Weinberger, J. (2014). Common factors are not so common and specific factors are not so specified: Toward an inclusive integration of psychotherapy research. *Psychotherapy*, *51*(4), 514–518. https://doi.org/10.1037/a0037092

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist*, *47*(12), 1578–1585. https://doi.org/10.1037/0003-066X.47.12.1578

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials. *Psychological Bulletin*, *130*(4), 631–663. https://doi.org/10.1037/0033-2909.130.4.631

Whiteford, H. A., Harris, M. G., McKeon, G., Baxter, A., Pennell, C., Barendregt, J. J., & Wang, J. (2013). Estimating remission from untreated major depression: A systematic review and meta-analysis. *Psychological Medicine*, *43*(8), 1569–1585. https://doi.org/10.1017/S0033291712001717

Widiger, T. A. (2012). *The Oxford Handbook of Personality Disorders*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199735013.001.0001

Woll, C. F. J., & Schönbrodt, F. D. (2020). A Series of Meta-Analytic Tests of the Efficacy of Long-Term Psychoanalytic Psychotherapy. *European Psychologist*, *25*(1), 51–72. https://doi.org/10.1027/1016-9040/a000385

Zanarini, M. C., Frankenburg, F. R., Reich, D. B., & Fitzmaurice, G. (2010). Time to Attainment of Recovery From Borderline Personality Disorder and Stability of Recovery: A 10-year Prospective Follow-Up Study. *American Journal of Psychiatry*, *167*(6), 663–667. https://doi.org/10.1176/appi.ajp.2009.09081130

Zilcha-Mano, S. (2019). Major developments in methods addressing for whom psychotherapy may work and why. *Psychotherapy Research*, *29*(6), 693–708. https://doi.org/10.1080/10503307.2018.1429691

I

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 20 May 2020
doi: 10.3389/fpsyt.2020.00384

Effectiveness of Open-Ended Psychotherapy Under Clinically Representative Conditions

Magnus Nordmo[1], Nils Martin Sønderland[1], Odd E. Havik[2], Dag-Erik Eilertsen[1], Jon T. Monsen[1] and Ole Andre Solbakken[1]*

[1] Department of Psychology, University of Oslo, Oslo, Norway, [2] Department of Clinical Psychology, University of Bergen, Bergen, Norway

Objective: This study investigates the effectiveness of open-ended psychotherapy in a large, naturalistic, and diverse patient cohort using rigorous and multifaceted assessments.

Method: Patients (N = 370) in open-ended psychotherapy completed an extensive set of self-report measures and diagnostic interviews, including long-term follow-up in order to assess stability of outcomes. About half of the patients qualified for a personality disorder at the onset of treatment. Treatments were open-ended, and on average therapists provided substantially larger treatment doses than common in the literature.

Results: A substantial majority recovered from their respective Axis I (58%) and/or Axis II (55%) disorders during treatment. Patients also experienced large positive changes in self-report measures of overall psychiatric symptoms and moderate positive changes in self-reported interpersonal problems, while very few (< 3%) demonstrated negative development. The patients maintained their diagnostic and self-assessed changes at a two-and-a-half-year follow-up. In contrast, self-reported occupational functioning showed minimal improvement throughout the treatment and follow-up phase.

Conclusion: A naturalistic patient cohort undergoing open-ended psychotherapy demonstrates substantial and stable improvements.

Keywords: psychotherapy, representative, effectiveness, outcome, naturalistic

OPEN ACCESS

Edited by:
Veena Kumari,
Brunel University London,
United Kingdom

Reviewed by:
Warren Mansell,
University of Manchester,
United Kingdom
William B. Stiles,
Miami University, United States

****Correspondence:***
Magnus Nordmo
magnus.nordmo@psykologi.uio.no

Specialty section:
This article was submitted to Psychological Therapies, a section of the journal Frontiers in Psychiatry

Received: *09 August 2019*
Accepted: *16 April 2020*
Published: *20 May 2020*

Citation:
Nordmo M, Sønderland NM, Havik OE, Eilertsen D-E, Monsen JT and Solbakken OA (2020) Effectiveness of Open-Ended Psychotherapy Under Clinically Representative Conditions. Front. Psychiatry 11:384. doi: 10.3389/fpsyt.2020.00384

INTRODUCTION

A considerable research effort has gone into the empirical investigation of the effects of psychotherapy. The goal of this research effort is to formulate accurate knowledge, which can inform health-policy and the practice of psychotherapy in general. One major challenge is to navigate the need for stringent protocols to increase internal validity on the one hand, with the need for "representative" research that can generalize to those institutions where the majority of psychotherapy takes place (1). Studies that seek to assess a treatment under real world conditions are typically labelled as effectivness trials, in contrast to lab driven efficacy trials. The distinction between efficacy and effectiveness research is somewhat nebulous as the two terms lack operationalized definitions. Prominent parameters separating the two include the selection of

participants, the use of treatment manuals, managing the "dose" of psychotherapy, and the type and amount of therapist training and monitoring (2). Researchers have demonstrated that these procedural variables can influence outcomes in a range of interventions (3). One dramatic example was when Weisz, Weiss & Donenberg (4) concluded that "most clinic studies have not shown significant effects" (p. 1578) after reviewing effectiveness psychotherapy interventions for children and adolescents. The authors contrasted this finding with evidence that efficacy trials from the lab consistently produced significant benefits. Also, what we consider "representative" is a moving target as in that the treatments delivered in routine practice change over time and across nationalities and regions.

There are arguably three main methodologies for assessing representative psychotherapy interventions for adults (5). One is to compare evidence-based treatments (EBT) with treatment as usual (TAU) condition (6, 7). If TAU is found to be equally effective, then this is indirect evidence that the TAU is itself effective. Wampold et al. (7) used meta-analysis to assess EBT *versus* TAU in 14 studies and found that EBT generally outperformed TAU, but that this difference was most likely an effect of heterogeneity in the TAU category, which also included minimal or no treatment. The superiority of EBT disappeared when the authors compared EBT with TAU conditions that contained psychotherapy (three studies). In contrast, Budge et al. (6) found that EBT was significantly more effective in treating personality disorders when compared to a TAU psychotherapy condition, although the amount and type of supervision, therapist training, and therapy dose were not balanced across the conditions (8).

The second line of research comes from direct benchmarking studies that make statistical comparisons between the results from efficacy and effectiveness trials (9). The evidence from benchmarking studies suggests that routine-care does indeed produce similar results compared to the lab (10–12). However, these results are most commonly from patient samples with an unknown degree of pathology with limited or no diagnostic assessments. Most studies only supply description of the primary symptomatic problem. This raises the question of whether patient characteristics such as severity of pathology or the presence of characterological problems might influence the relationship between trial representativeness and outcome. Lastly, the majority of data from benchmarking trials utilize treatment data from American university counseling centers (13). Thusly, these results might not generalize to non-university clinical settings.

The third line of research uses meta-analytical tools to investigate representativeness. Shadish et al. (14) categorized studies with a criterion-based approach. To pass the first criterion the studies had to be conducted in a non-university setting, with patients that were referred *via* traditional routes and used professionals with regular caseloads. To pass the second criterion the studies had to pass criterion one and not use a treatment manual or any monitoring of intervention implementation. To pass the third criterion the study had to pass criterion two, use clients that were heterogeneous with respect to presenting problems, personal characteristics and lastly, not include any explicit therapist training immediately before the study. From 1,082 possible studies that the authors included in the initial pool of meta-analyses, only 56 (5.2%) passed criterion one, 15 (1.4%) studies, passed criterion two and only one study (0.1%) passed criterion three. That study was a family therapy intervention for children (n = 11) with behavioral problems (15). Shadish et al. (14) concluded that criterion one-studies appear to produce similar effect sizes compared with the total sample, but that the lack of criterion two and three studies prohibits strong conclusions regarding the effect of routine-care *versus* lab trials. In a follow-up study (3), the authors expanded the base pool of studies and refined the criteria for representativeness. In place of the earlier stage system, the authors took a dimensional approach, coding the degree of representativeness on a scale from zero to ten. With an expanded pool of baseline studies, including a set of highly representative "clinic therapy" (p. 513) studies, as well as a non-representative set of randomized controlled trials, the authors found that the effects of psychotherapy were robust across the spectrum. A regression analysis of the clinical representativeness features indicated that a large therapy dose, the presence of an internal control group, homogeneity in presenting problem, use of structured therapy, and flexibility in the number of sessions were positively related to effect sizes.

In more recent years, many large-scale randomized pragmatic trials have been conducted, where the aim is to assess the effectiveness of a specific EBT in routine care. These generally support the effectiveness of EBT implemented in routine care (5). A predicament of the pragmatic trial is that it has to balance the need for external validity with the necessity of treatment fidelity. In practice, this means selecting or training therapist to deliver a particular treatment. The very features that the pragmatic trials seek to achieve, namely experimental control, make generalizations to non-EBT clinical settings precarious, although less so compared to the classical lab-driven RCT. The problem of generalization is highlighted by evidence suggesting that therapists seldom implement EBT in their routine care (16). On a related note, evidence from psychopharmacological research suggests that effectiveness trials demonstrate lower effects when compared to their experimental RCT counterparts (17).

Based on this summary of psychotherapy studies under representative conditions it seems that the majority of evidence comes from either a synthesis of heterogenic meta-analytic investigations or from pragmatic trials and benchmarking trials that assess a specific EBT. This synthesis suggests that psychotherapy, on the whole, is effective when implemented under routine-care conditions, but when limiting analyses to highly representative studies, the evidence is limited. Studies have typically severely restricted the number of sessions and rarely provided diagnostic information beyond symptomatic assessment. These characteristics constitute a challenge as the severity of psychopathology, and the presence of personality disorder represent confounding variables. Similarly, "real-world" effectiveness investigations are almost exclusively from university counseling centers. Our goal in this study is to assess the

effectiveness of open-ended psychotherapy in a representative healthcare setting with a large, heterogeneous sample, including severe characterological psychopathology. We believe that these results may serve to supplement a literature that is dominated by milder varieties of mental problems. Assessments include rigorous diagnostic interviews as well as patient's self-assessments, both with a considerable follow-up period beyond treatment termination. To our knowledge, the effectiveness of psychotherapy with a representative patient sample, which does not asses a particular EBT, has not been previously documented in the psychotherapy literature. In the present paper we answer the following research questions:

Given an open-ended, naturalistic, and representative psychotherapy setting,

1. What are the rates and magnitudes of diagnostic, symptomatic, and interpersonal change?
2. To what degree are therapeutic gains maintained over time?
3. Do patients experience a positive change in occupational status?

METHODS

Study Overview

We adopted treatment data collected in the Norwegian Multicenter Study of Process and Outcomes in Psychotherapy (NMSPOP). The NMSPOP is a naturalistic study with a total sample of outpatients (N = 370) gathered from eight treatment sites within the Norwegian public health system in the years 1995–2008. The majority of patients (n = 301) were recruited from psychiatric outpatient clinics spread across 17 separate Norwegian clinics. We also gathered data from the Norwegian University of Science and Technology`s student clinic (patient n = 27). Lastly, we gathered data from outpatient clinics with physiotherapists (patient n = 42) undergoing specialization in psychodynamic body therapy for patients with somatoform disorders (18).

At each of the eight sites, trained coordinators (clinical psychologist or psychiatrist) were responsible for recruiting patients and administering the research protocol. We instructed the coordinators to select patients from their local population randomly, but also to ensure that roughly half had a diagnosable personality disorder. We did not apply any formal randomization procedure. The local coordinators also assessed the patients. The coordinators were all experienced clinicians who underwent training using the assessment instruments. The inclusion policy was liberal, with the following exclusion criteria: age less than 20 years, active psychosis, drug/alcohol abuse as the primary problem, need for emergency treatment or hospitalization, and mental retardation (IQ < 70). These criteria are in line with commonly used criteria in the evaluation of patients for individual psychotherapy at outpatient clinics. The Regional Committee for Medical Research Ethics in Eastern Norway approved the study.

After receiving information and signing a written consent, the patients were submitted to a two-step pretreatment assessment. In the first step, patients completed several self-report questionnaires, including among others a sociodemographic inventory, occupational functioning, the Symptom Checklist-90-Revised [SCL-90-R: (19)] and the Inventory of Interpersonal Problems 64 [IIP 64: (20)]. In the second step, patients underwent a structured diagnostic assessment by the coordinator at each site. This assessment comprised of a Structured Clinical Interview (SCID) based on the Diagnostic and Statistical Manual of Mental Disorders 4th edition (21) criteria for Axis I and II disorders. All assessment interviews and therapy sessions were audio recorded. A subset of the SCID I and SCID II interviews were blindly double-coded by an independent professional to assess inter-rater reliability. The patients were assigned to therapists based on availability after the initial assessment. Patients completed self-report questionnaires during treatment after the 3rd, 12th, and 20th session. Patients completed self-report questionnaires every 20th session following the 20th session for as long as they received therapy. Following treatment completion, the coordinator repeated the diagnostic evaluation with a SCID I and II interview. The self-report questionnaires were also repeated at the posttreatment assessment. While some of the patients completed their postassessments directly after treatment completion (35%), most completed this assessment a few months after treatment completion due to practical issues. The average delay was 9.8 after treatment completion (SD = 25.5, Median = 2.46). The patients were then assessed with SCID interviews by the same coordinator and completed self-report measures six months, one, and two and a half years following the posttreatment assessment. A subsample (n = 17) of patients also had a six year follow-up assessment.

The therapists (n = 88) were mainly experienced clinicians with a mean of 10 years (SD = 6.5) of psychotherapy experience. All therapist also had postgraduate professional training, including a mean of 5.9 years (SD = 4.3) of clinical supervision. Notable exceptions were the physiotherapists (n = 8) and student therapists (n = 27) who received supervision. The mean number of patients per therapist was 5.6 excluding the student therapists where each student saw one patient. Therapists were instructed to provide their usual therapeutic practice.

Sample Characteristics

The mean sample age was 35.2 (SD = 9.4) years with a majority of female patients (69.5%). See **Table 1** for a description of pretreatment sample diagnostic status. The mean number of pretreatment SCID II criteria of personality disorder was 12.7 (SD = 5.8). The pretreatment mean symptom score, as measured by the Global Severity Index (GSI), was 1.28 (SD = 0.61), while the mean rating of interpersonal problems (IIP Global) was 1.49 (SD = 0.52). The patients reported that "the problem which you are now seeking treatment for" had lasted on average 11.7 years (SD = 9.75). The sample did not include patients who sought treatment for a primary substance use diagnosis. The Norwegian healthcare system has a separate subdivision with clinics specializing in the treatment of primary substance use disorders.

When assessed pretreatment, a subgroup of patients indicated that they used prescribed medication to treat their psychological

TABLE 1 | Changes in occupational status and diagnosis frequency.

	Number and percentages of patients		
	Pretreatment*	**End of treatment****	**Two-year follow-up*****
Functioning	202 (55 %)	205 (65 %)	190 (65 %)
Non-functioning	154 (42 %)	110 (35 %)	104 (35 %)
SCID1 Diagnosis			
Presence of SCID 1 diagnosis	321 (87 %)	119 (40 %)	99 (38 %)
Affective disorders	150	56	42
Anxiety disorders	406	109	84
Somatoform disorders	117	26	23
Eating disorders	31	10	4
Substance-Related disorder	9	4	4
Schizophrenia and other psychotic related disorders	2	3	2
SCID2 Diagnosis			
Presence of SCID 2 diagnosis	200 (54 %)	84 (28 %)	58 (22 %)
Cluster A	82	31	14
Cluster B	71	24	21
Cluster C	169	62	35
Not Otherwise Specified	2	0	1

Percentage from total sample (N = 370).
***Percentage from available occupational (n = 315) and diagnostic (n = 297) post-treatment data.*
****Percentage from available occupational (n = 294) and diagnostic (n = 258) follow-up data.*

problems either "regularly" (22%) or "when in need" (7%). The majority of patients using psychotropic medication indicated that they mainly used an antidepressant (n = 72), while fewer indicated that they mainly used an anxiolytic (n = 19), a hypnotic (n = 3), an antipsychotic (n = 4), or pain medication (n = 8). Of the psychotropic medication users, the majority used a single medication (n = 70), while some were prescribed two (n = 22), three (n = 9) or four (n = 4) different medications.

Assessment Instruments

The Symptom Checklist 90 Revised (SLC-90)

We used the SLC-90 to assess overall symptom presence and severity. It contains 90 questions asking patients to rate, on a Likert scale from 0 (not at all) to 4 (very much), the intensity of a given symptom during the last week. The symptoms represent nine dimensions of distress, which can be further grouped into three global indexes (22). We used the Global Severity Index (GSI) as an overall symptom severity measure, which is the mean rating across the entire checklist. The GSI is a robust measure of overall symptom severity (23). The SCL-90 demonstrated high internal validity in our sample with a pretreatment Cronbach's alpha of.97.

The Inventory of Interpersonal Problems 64 (IIP-64)

We applied the IIP-64 to assess levels of interpersonal problems. It consists of 64 questions rated on a five-point Likert scale from 0 (*not at all*) to 4 (*very much*). The first 39 questions begin with the phrase "*It is hard for me to…*" while the remaining 25 questions ask about "*Things that I do too much.*" We used the IIP global, which is the mean scores across the entire inventory. This global score has been shown to adequately capture a wide range of interpersonal problems and pathology (24). The IIP-64 demonstrated high internal validity with a pretreatment Cronbach's alpha of.93.

Occupational Status

We created a dichotomous variable (occupational functioning vs. no functioning) based on self-reported occupational status. We classified the following responses as "functioning": 1) "I am currently engaged in paid work," 2) "I am a stay-at-home mom/dad," 3) "I am currently engaged as a student" or 4) "I am retired." We classified "non-functioning" with the following responses: 1) "I am currently on sick leave," 2) "I am currently in work rehabilitation," 3) "I am currently receiving disability benefit" or 4) "I am currently unemployed." We assessed changes in occupational status by comparing pre- to posttreatment levels and pre- to follow-up measurements. We used the last recorded assessment in our follow up assessment which was six years for 4.9% (n = 17) of the sample, two and a half years for 68.6% (n = 254), one year for 5.1% (n = 19) and six months for 1.0% (n = 4). A total of 20.5% (n = 76) and 14.9% (n = 55) had no follow-up measurement of occupational status at follow-up and posttreatment respectively.

SCID Interview

The SCID interview was developed for the assessment of both Axis I clinical disorders (SCID I) and Axis II personality disorders (SCID II) according to the Diagnostic and Statistical Manual of Mental Disorders (21). Each disorder is associated with a set of items that assess different manifestations of the disorder. Each item can be scored as either absent, sub-threshold, true, or "inadequate information to code." The items corresponding to each particular disorder is then summed to assess whether the patient qualifies for a given disorder. The SCID interviews have been shown to give reliable assessments of DSM-IV diagnoses (25). To assess changes in diagnostic status for each Axis, we employ a dichotomy between no diagnosis (0) and one or more diagnoses present (1).

A sample of 40 SCID I and 20 SCID II interviews were selected at random to assess inter-rater reliability. We found that Cohen's Kappa for Axis I disorders ranged from.53 to 1.00 with a mean of.75, indicating fair too excellent agreement. Cohen's Kappa for Axis II disorders ranged from.63 to 1.00 with a mean of.82, indicating good to excellent agreement.

Statistical Analysis

We carried out all statistical analyses using R version 3.5.2 (26). We used the lme4 package (27) to fit linear mixed models, and lmerTest (28) equipped with Satterthwaite's degrees of freedom method for p-values. We used ggplot2 (29) and ggalluvial (30) to make the figures. Imputation was performed using the mice package (31).

Longitudinal Self-Report Questionnaires

We analyzed the data using Multilevel Modeling (MLM) with Bayesian Information Criteria (BIC) as our indicator of model fit. MLM is the recommended method for analyzing longitudinal and repeated health measures as it allows the nesting of each

measurement (level 1) within each patient (level 2). MLM has been shown to outperform traditional methods (*e.g.*, last observation carried forward method) when handling missing data and accounting for potential dropout bias (32, 33). For our treatment phase model, we applied the session number as a fixed occasion time estimate, as this allows for between-subjects comparisons of regression coefficients. We centered treatment start at zero and coded the posttreatment assessment as the last session number from the longest treatment series, plus 1, which was 361. For our follow-up model, we centered time as zero at treatment completion and used month after treatment completion as our time measure.

We began our analysis by visually inspecting GSI and IIP-Global raw scores. We observed a log-linear distribution for the majority of cases on both outcome measures, both in the treatment- and follow up phase. To assess this, we ran models with both linear time, where each time-point corresponds to the specific time of measurement, and with log-transformed time, where each time-point is multiplied by the $\log^{10}$ to produce a log-linear curve. These analyses confirmed the superior fit of log-linear slopes on both our main self-report outcome measures with lower BIC values as compared to linear time models. The superior fit of log-time was true for both the treatment phase and the follow-up phase, although the rate of change was marginal during follow-up. Another benefit that is of particular importance in our open-ended design is that log-linear time places the last observation (*e.g.*, the postmeasurement) closer to its original time-point, thereby lessening any potential skew produced coding post time as 361 on patients with short treatments (34).

Several investigations have shown that latent therapist effects can influence the statistical modeling of patient trajectories (35–37). Therefore, in addition to the analyses presented below, we also carried out a separate three-level (measurements, patients, therapist) model. This model produced similar overall results and a poorer BIC value compared to the two-level model (measurements, patients). We also performed a separate analysis with patients nested within treatment site, also with very similar results and worse BIC value, leading us to omit these results.

Effect Sizes and Clinically Significant Change

We used Cohen's *d* to estimate the magnitude of change by dividing the estimated change score with the corresponding pooled standard deviation. The pooled standard deviation was estimated by taking the mean standard deviation from all measurement points with more than 150 patients, and merging the treatment phase and follow-up phase into a single measure. The pooled standard deviations were 0.56 for GSI and 0.51 for IIP Global. Our post measurement standard deviations were counted twice as both the end of the treatment phase, and as the start of the follow-up phase, corresponding to our two multi-level models. Using a pooled standard deviation reduces the problem of artificially inflating the effect size (38). We applied Cohen's (39) standard for categorizing effect sizes, where 0.2–0.5 is a small effect size, 0.5–0.8 is a moderate effect size, and > .08 is a large effect size. Clinically significant change was calculated according to Jacobsen and Truax's (40) definition in which a patient's predicted score needs to achieve both a reliable statistical change (41) and pass the cut-off to a functional population as compared to a dysfunctional population. We used the Norwegian norms from Carrozzino et al. (42) and Monsen, Hagtvet, Havik, and Eilertsen (43) and calculated these cut-offs to be 0.73 and 1.22 for the GSI and IIP Global, respectively. We defined a patient as recovered if he or she met both criteria. We further categorized patients as reliably improved if they demonstrated a statistically reliable change, but failed to cross over the cut-off of dysfunction; unchanged, if they did not demonstrate a statistically reliable change; or deteriorated, if a patient showed a negative, statistically reliable change. We calculated these criteria for both of our main outcome measures at the end of treatment and the final follow-up assessment.

Diagnostic and Occupational Status

Using the exact2x2 R package (44), we applied two-sided McNemar tests with continuity corrections and odds-ratio tests to analyze changes in diagnostic and occupational status from pre- to posttreatment, and pretreatment to follow-up status. This test obtains its respective p-values through central hypergeometric distribution (45). Both measures were coded as either 0 (No diagnosis/Functioning) or 1 (Diagnosis/Non-functioning).

Registration

To increase transparency, we registered our hypotheses and statistical analyses using the Open Science Framework (46) before running our analyses. We completed this registration after the data had been collected and preliminary analyses had been conducted. The results of the preliminary analyses were presented at the 39th International Meeting of the Society for Psychotherapy Research conference. We have made deliberate efforts to prevent the preliminary analyses from affecting our current results by not subdividing our data, including all relevant outcomes measurements and selecting standard statistical tools for multilevel longitudinal data.

RESULTS

Patient Flow and Data Completeness

Out of the original sample of 370 patients, eight did not start treatment or did not give any assessment following the pretreatment assessment, giving a total treatment sample of 362. When analyzing diagnostic and occupational follow-up data, the last available assessment was used. A few patients had a last measurement of six years (n = 23) after treatment, while the majority of the patients had follow-up assessments approximately two and a half years after treatment completion (Mean = 28.1 months, SD = 8.6). A total of 52 (14.1%) patients had no follow-up measurement of main outcomes. Logistical regression was performed on both the GSI and IIP using pre- and posttreatment scores to predict missing follow-up data. Patients' pre (p = 0.79) and post (p = 0.55) treatment GSI scores did not predict missing GSI at follow-up. However, both pre-, $\chi 2(1)$ =

1.32, $p = 0.014$, and posttreatment, $\chi2(1) = -1.13$, $p = 0.027$ IIP Global scores were statistically significant predictors for missing IIP Global at follow-up. The multi-level models were fitted with all available data. The data collection of self-report, diagnostic and occupational status represents distinct and separate processes and therefore have distinct attrition patterns. See **Figure 1** for a complete description of patient attrition.

The distribution of therapy length is shown in **Figure 2**. The mean number of sessions was 51.3 (SD = 58.9) with a median of 35. A few patients with very long therapies accounted for the high variance. Every treatment was terminated by a joint agreement between the therapist and the patient except for therapy dropouts with one exception: One of the sites (therapist n = 7, patient n = 31) had an upper-limit of 40 sessions.

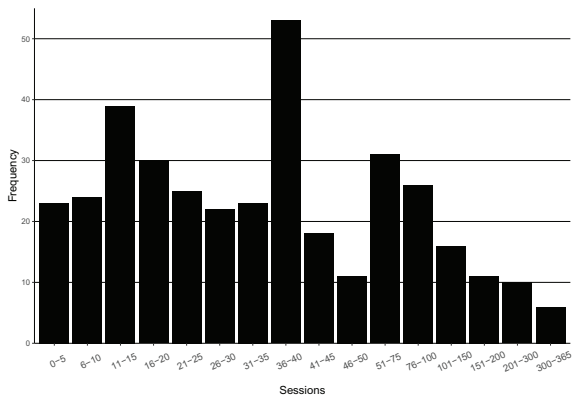


FIGURE 2 | Treatment length.

Diagnostic and Occupational Status

Table 1 shows changes in diagnostic and occupational status. We apply the terminology that positive change equals fewer diagnoses. We found a positive statistically significant difference of any symptom diagnosis between pre- to posttreatment, $x^2(1, n = 293) = 128.8$, $p < 0.01$, OR = 19.1, 95% CIs [9.5, 45.1]. We also found a positive statistically significant change when comparing presence of any symptom diagnosis pretreatment to follow-up, $x^2(1, n = 254) = 109.5$, $p < .01$, OR = 15.1, 95% CIs [7.7, 33.8]. When applying the same test with the presence of any personality disorder diagnosis we found the positive changes from pre- to posttreatment to be statistically significant, $x^2(1, n = 294) = 59.2$, $p < .01$, OR = 8.1, 95% CIs [4.3, 16.8]. Lastly we found a positive statistically significant change in presence of any personality disorder from pretreatment to follow-up status, $x^2(1, n = 257) = 65.7$, $p < .01$, OR = 8.2, 95% CIs [4.5, 16.3].

We found that overall, changes in self-reported occupational status were minimal through the treatment and follow-up phase. The majority of patients kept their respective status through the treatment and the follow-up phase, while a substantial minority went from non-functioning to functioning. Relatively few patients had a negative development from functioning to non-functioning. See **Figure 3** for an alluvial development diagram. A McNemar test revealed a statistically significant positive change in occupational functioning from pre- to posttreatment, $x^2(1, n = 310) = 5.8$, $p = 0.12$, OR = 1.68, 95% CIs [1.1, 2.6], indicating that more people were able to work posttreatment, compared to pretreatment. The trend was maintained when we compared pretreatment to follow-up occupational status, $x^2(1, n = 291) = 2.4$, $p = .12$, OR = 1.38, 95% CIs [0.9, 2.1].

To account for missing occupational data we performed a pair of analysis using the Multivariate Imputation by Chained Equation methodology. Missing occupational data were imputed by using available diagnostic and occupational status, gender, GSI, and IIP Global. We created 30 datasets with a maximum of 20 iterations, using parallel socket cluster. When analyzing the complete dataset we found that pre- to posttreatment remained significant, $x^2(1, n^{imp} = 370) = 6.7$, p = 0.02, OR = 1.56, 95% CIs [1.1, 2.3], as did pre- to follow-up, $x^2(1, n^{imp} = 370) = 5.48$, p =.02, OR = 1.51, 95% CIs [1.1,

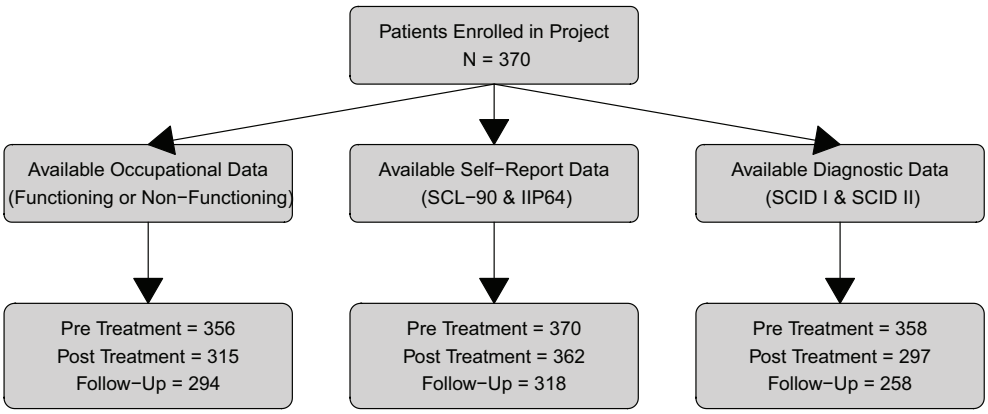


FIGURE 1 | Patient attrition.

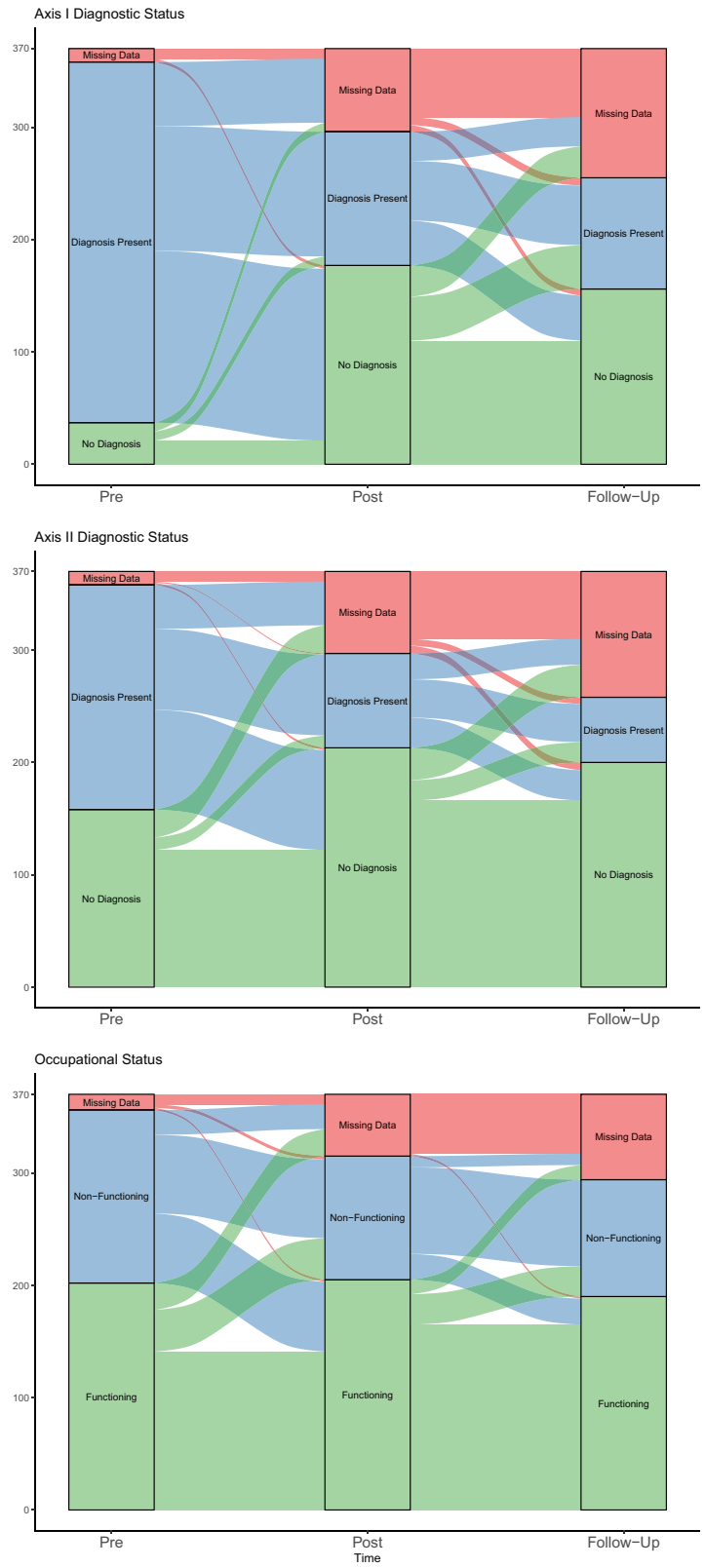


FIGURE 3 | Alluvial diagram of changes in diagnostic and occupational status.

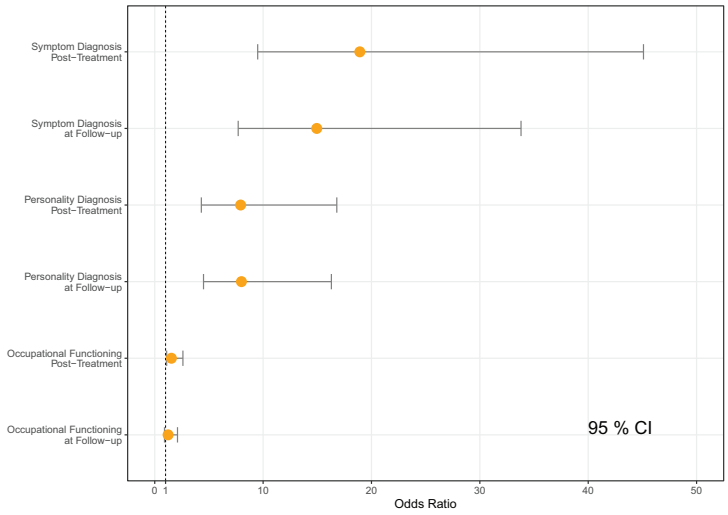


FIGURE 4 | Odds ratios comparison to pretreatment.

2.1]. See **Figure 4** for a graphical visualization of Odds Ratio scores.

Change in Symptoms and Interpersonal Problems During Treatment

The multilevel models from the main outcome treatment phase are shown in **Table 2**. We found that scores on the GSI were subject to a statistically significant change during the treatment phase when allowing for variable change across patients (Model 1), $\beta = -.18$, $p < .01$. This was also the case for IIP Global (Model 1), $\beta = -.11$, $p < .01$. For the GSI, the overall intercept in the treatment phase was estimated to be 1.32. Overall change across the treatment phase was estimated to be a reduction of.47 points. For IIP Global the overall intercept was estimated to be 1.45 with a total treatment change of 0.29. There was substantial heterogeneity across the sample concerning both the GSI and IIP Global as indicated by the significant variance in the intercepts. The significant variance in slopes on both main measures indicates a high degree of variability in treatment response, which emphasizes the need for multilevel analysis. The necessity of multilevel analysis is also indicated by the lower BIC when comparing Model 1, which allows for variable change across patients, with Model 0, which does not. See **Figure 5** for a visualization of predicted scores.

TABLE 2 | Measures on symptoms and interpersonal functioning in the treatment phase.

	GSI		IIP Global	
	Model 0 Est	**Model 1 Est**	**Model 0 Est**	**Model 1 Est**
Fixed effects				
Intercept	1.33** (.03)	1.33** (.04)	1.45** (.03)	1.46** (.03)
Logtime	-.186** (.01)	-.186** (.01)	-.108** (.008)	-.113** (.012)
Random effects	Est	Est	Est	Est
Residual	.14** (.37)	.10** (.32)	.09** (.30)	.07** (.26)
Variance in intercept	.31** (.56)	.39** (.63)	.24** (.49)	.027** (.02)
Variance in slopes	N/A	.04** (.21)	N/A	.03** (.17)
Intercept Slope Corr.	N/A	-.43	N/A	-.32
BIC	2714.5	2555.8	1952.8	1798.3

*Standard error in parenthesis. Estimation performed using Restricted Maximum Likelihood (REML). Model 0 fixates the rate of change for each patient while Model 1 allows for variable changes across patients. Fixed effects are presented with the estimate and standard deviation in parenthesis. Random effects are presented with the variance and standard error in parenthesis. **p < .01.*

Symptoms and Interpersonal Problems in the Follow-Up Phase

The results from the multilevel model of both main outcome variables during the follow-up phase are shown in **Table 3**. The follow-up phase was associated with a small but significant drop in interpersonal problems as measured by the IIP Global when allowing for variable changes across patients, $\beta = -.04$, $p < .001$. The overall intercept was estimated to be 1.19 with a total reduction of.064. The GSI also show a further reduction, but this change was not significant $\beta = -.008$, $p = 0.49$. The overall intercept in the follow-up phase was estimated to be 0.81 with a total reduction of.012 points. Overall, this indicates that treatment gains were maintained. The follow-up phase also showed significant variance in both intercepts and slopes.

Clinically Significant Change and Effect Sizes

Table 4 shows clinically significant change at the level of the individual (40). The majority of patients (69%) experienced a reliable improvement or complete recovery in terms of overall psychiatric symptoms as measured by the GSI. This symptomatic improvement was sustained during the follow-up period. For interpersonal problems, about one third (35%) of the patients made either complete recovery or a reliable positive change. The

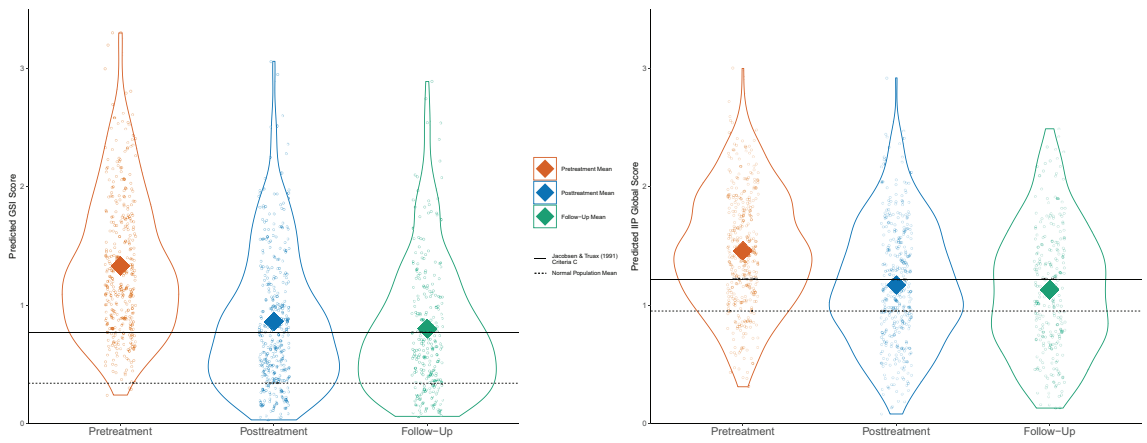


FIGURE 5 | GSI and IIP Global predicted score distributions.

IIP Global revealed fewer complete recoveries compared to the GSI. During the treatment phase, the GSI change was large (d = 0.85) while the IIP Global change was moderate (d = 0.57). The follow-up phase showed high stability, with the GSI demonstrating a weak improvement (d =.03) which was not statistically significant, while the IIP Global demonstrated a somewhat stronger improvement (d = 0.13) which was statistically significant.

DISCUSSION

This study shows that patients who receive psychotherapeutic care in an open-ended outpatient format, experience large to moderate positive change on self-report measures of overall psychiatric symptoms and interpersonal difficulties, as well as large observer-rated diagnostic improvements. The majority of patients do not qualify for an Axis I or Axis II diagnosis at the end of treatment and can be categorized as either reliably improved or recovered. Patients experienced a more substantial reduction in general psychiatric symptom compared to interpersonal problems which they several years after treatment termination. During the follow-up phase, the patient's experiences further positive changes in interpersonal functioning, whereas they maintained their level of general psychiatric symptoms. The improvements seen in interpersonal functioning during the follow-up might be a positive consequence of the open-ended nature of treatment, where patients and therapists are free to focus on characterological problems in contrast to manualized, time-limited and symptom-focused interventions.

As this study is a cohort study without a control group, we cannot ascertain a causal link between patient improvement and the psychotherapy received. Due to ethical concerns (47) and practical challenges (48), few studies focus on untreated

TABLE 3 | Measures on symptoms and interpersonal functioning in the follow-up phase.

	GSI		IIP Global	
	Model 0 Est	**Model 1 Est**	**Model 0 Est**	**Model 1 Est**
Fixed effects				
Intercept	.82** (.04)	.82** (.04)	1.19** (.03)	1.19** (.03)
Logtime	-.009 (.01)	-.008 (.02)	-.042** (.01)	-.043* (.01)
Random effects	Est	Est	Est	Est
Residual	.08** (.28)	.07** (.26)	.07** (.26)	.06** (.23)
Variance in intercept	.34** (.58)	.38** (.61)	.28** (.53)	.30** (.55)
Variance in slopes	N/A	.03* (.17)	N/A	.02** (.14)
Intercept Slope Corr.	N/A	-.32	N/A	-.22
BIC	1199.8	1191.7	996.9	995.8

*Standard error in parenthesis. Estimation performed using Restricted Maximum Likelihood (REML). Model 0 fixates the rate of change for each patient while Model 1 allows for variable changes across patients. Fixed effects are presented with the estimate and standard deviation in parenthesis. Random effects are presented with the variance and standard error in parenthesis. *p < .05 **p < .01.*

TABLE 4 | Clinically significant change using predicted scores.

	Number and Percentages of Patients	
Measures and status	**Pretreatment to termination of treatment ***	**Pretreatment to two-year follow-up ****
Global Severity Index (SCL-90)		
Recovered	142 (38 %)	132 (42 %)
Improved	114 (31%)	72 (23 %)
Unchanged	107 (29 %)	103 (32 %)
Deteriorated	6 (1.6 %)	11 (3.4 %)
IIP Global (IIP-64)		
Recovered	84 (23 %)	99 (31 %)
Improved	43 (12 %)	42 (13 %)
Unchanged	234 (63 %)	167 (53 %)
Deteriorated	9 (2.4 %)	10 (3.1 %)

** Percentage from total sample (N = 370).*
*** Percentage from patients with follow-up measure (n = 318).*

psychiatric populations. The lack of studies on untreated psychiatric populations makes it difficult to compare our treatment sample with a hypothetical no-treatment control. The studies that do exist are mainly on patients with a mild to moderate mood and/or an anxiety disorders (49, 50). These studies suggest that spontaneous recovery for anxiety and mood disorders is common, but estimates vary greatly between studies. Spontaneous recovery is documented to be rarer for patients diagnosed with a comorbid personality disorder (51–54) than for patients without. There is a lack of data for recovery rates of severely ill populations that are not in treatment as these individuals are typically high-treatment utilizers (55). We would argue that the samples presented in the spontaneous recovery literature is of milder psychopathology compared to our sample, where 54% fulfilled the criteria for one or more personality disorders at admission. The robust positive changes seen in our sample seems greater than what one might expect from the rate of natural recovery found in each disorder. We also believe that the observation that our sample maintained the therapeutic gains during the follow-up is indicative that positive changes should be ascribed the therapy received. This lasting change is contrasted to the chronicity reported pretreatment. Although randomization to a control condition can open the road to causal analyses, we believe that that this is precarious for research on long-term treatments for severely ill patients that seek to compare routine care and spontaneous recovery. In such a scenario, the control condition would have to limit or omit routine care for several years forcefully. This dynamic is also apparent in the controversy surrounding evidence-based therapies (56).

The substantial positive changes in overall symptoms and interpersonal problems did not correspond to an equally substantial positive change in occupational functioning. The changes seen in occupational functioning were negligible. This finding is sharply contrasted by the clinical recovery and improvements observed for the majority of patients. This finding is in line with research that indicates that disability interventions that exclusively focuses on the treatment of a mental disorder rarely produces occupational recovery (57) as they overlook the complex interaction of work perceptions and challenges, attitudes, beliefs and other psycho-social influences (58, 59). Our findings indicate that patients can improve substantially, as measured by self-reported mental health questionnaires and observer-rated psychiatric diagnosis, and still see a meager degree of positive change in occupational status. This result should be interpreted with caution as missing posttreatment and follow-up data obscures the analysis. Another concern is related to our measurement of occupational functioning. Returning to work is a complex phenomenon (60) and a single-item operationalization might lack the sensitivity to detect intricate changes.

The outcome measure with the most substantial overall change was psychiatric diagnoses as measured by the SCID interviews. The majority of patients did not qualify for a diagnosis after the treatment phase. This improvement was largely maintained in the follow-up phase and true for both Axis I and II diagnoses. Previous research has found that clinicians and independent observers usually report more substantial positive change when compared to patients assessing themselves (61, 62). In the present study, an independent coordinator, and not the respective clinician performed the diagnostic interviews and assessment, so the comparison to research on clinician versus patient ratings may be inaccurate. We believe that the independence of the coordinators adds to both the validity and reliability of our diagnostic data.

Using predicted scores, we found a surprisingly low amount of deterioration (1–3%) compared to what is generally found in the adult psychotherapy (5–10%) literature (5). We did not expect low amounts of deterioration as effectiveness interventions are usually associated with higher rates of deterioration when compared to structured efficacy trials (63). We believe that the low rate of deterioration is caused by either of two explanations or a combination of the two: Firstly, the fact that this is an open-ended treatment might have given the therapist the possibility of sustaining treatment until he or she was convinced that the patient was well enough to terminate treatment. This feature is in contrast to treatments with a prescribed set of sessions where a patient might face a setback or drop in functioning at the end of the prescribed amount of sessions, thereby giving the patient a negative skew on his or her predicted slope. Indeed, others have argued that time constraints can have an impact on successful termination (64). With an open-ended format, the patient can continue in therapy until the crises have been overcome and normal functioning has been established, thereby lessening the negative skew effect of the transient crises. Indeed, it has been posited that the experience of control over therapy conditions is directly intertwined with patient improvement (65).

Limitations and Future Directions

A few limitations should be considered when assessing the degree of representativeness of our procedures and sample. Firstly, we did not randomize the selection of therapist to participate in this trial; rather therapists were self-recruited based on availability at each site. Second, patients were not formally randomized to participate in the trial. Instead, the local administrator of each treatment site was instructed to pick out patients as randomly as possible, while striving to provide a representative sample for his or her respective treatment site. No information was collected from patients who declined the invitation to participate. Our therapist sample was comprised mostly of experienced professionals who volunteered to be a part of an intensive research project. A potential selection effect could have excluded underperforming therapist who might have contributed to a higher rate of deterioration (66).

Another, more substantial question is to what degree this study captures the elements of what is considered "representative" psychotherapy practice. This question is hard to answer as routine care is a moving target that changes with time and across national and regional borders. Examples of this process can be seen

in the proposed movement towards evidence-based and stepped-care treatments initiated with the Improved Access To Psychological Therapies (IAPT) reform in the UK (67) and its American (68) and Norwegian (69) equivalent. These are examples of possible healthcare reform that could substantially change what we mean by psychotherapy in routine care. The moving target phenomenon is also evident from the psychotherapy research literature. Virtually no psychotherapy effectiveness research has been published in the last ten years when defining effectiveness as routine care that does *not* prescribe a specific treatment for a specific diagnostic population. Therefore, we believe that our results fill an important missing piece in the current psychotherapy literature. In our opinion, these results make a fair representation of what one could expect from the "classical" way of delivering psychotherapy, where therapists are free to choose the treatment methodology, do not have a preset time-constraint, and treat each patient individually, face-to-face. We plan on completing further moderator analyses in future publications.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Regional Committees for Medical Research Ethics —South East Norway. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The work presented is the result of a cooperative effort from the respective authors. Each author has made significant contributions in the development of this manuscript. This includes, but is not limited too, planning the publication, writing the manuscript, analysing data and formulating theoretical and clinical implications.

FUNDING

The present article was funded by grants from: 1. The Norwegian Research Council; Health and Rehabilitation through the Norwegian Council for Mental. 2. The Department of Psychology, University of Oslo.

REFERENCES

1. Nathan PE, Stuart SP, Dolan SL. Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *psychol Bull* (2000) 126:964–81. doi: 10.1037/0033-2909.126.6.964
2. Seligman MEP. The effectiveness of psychotherapy: The Consumer Reports study. *Am Psychol* (1995) 50:965–74. doi: 10.1037/0003-066X.50.12.965
3. Shadish WR, Navarro AM, Matt GE, Phillips G. The effects of psychological therapies under clinically representative conditions: A meta-analysis. *psychol Bull* (2000) 126:512–29. doi: 10.1037/0033-2909.126.4.512
4. Weisz JR, Weiss B, Donenberg GR. The lab versus the clinic: Effects of child and adolescent psychotherapy. *Am Psychol* (1992) 47:1578–85. doi: 10.1037/0003-066X.47.12.1578
5. Lambert MJ. The efficacy and effectivness of Psychotherapy. In: Lambert MJ, editor. *Bergin and Garfield"s Handbook of Psychotherapy and Behavior Change*, *6th edition*. N.J. Hoboken: John Wiley & Sons (2013). p. 93–139.
6. Budge SL, Moore JT, Del Re AC, Wampold BE, Baardseth TP, Nienhuis JB. The effectiveness of evidence-based treatments for personality disorders when comparing treatment-as-usual and bona fide treatments. *Clin Psychol Rev* (2013) 33:1057–66. doi: 10.1016/j.cpr.2013.08.003
7. Wampold BE, Budge SL, Laska KM, Del Re AC, Baardseth TP, Flűckiger C, et al. Evidence-based treatments for depression and anxiety versus treatment-as-usual: A meta-analysis of direct comparisons. *Clin Psychol Rev* (2011) 31:1304–12. doi: 10.1016/j.cpr.2011.07.012
8. Wampold BE, Imel ZE. *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work*. New York, NY: Routledge (2015).
9. Hunsley J, Lee CM. Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. *Prof Psychol: Res Pract* (2007) 38:21–33. doi: 10.1037/0735-7028.38.1.21
10. Minami T, Wampold BE, Serlin RC, Hamilton EG, Brown GS, Kircher JC. Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *J Consult Clin Psychol* (2008) 76:116–24. doi: 10.1037/0022-006X.76.1.116
11. Minami T, Davies DR, Tierney SC, Bettmann JE, McAward SM, Averill LA, et al. Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *J Couns Psychol* (2009) 56:309–20. doi: 10.1037/a0015398
12. Wade WA, Treat TA, Stuart GL. Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *J Consult Clin Psychol* (1998) 66:231–9. doi: 10.1037/0022-006X.66.2.231
13. McAleavey A, Youn SJ, Xiao H, Castonguay LG, Hayes JA, Locke BD. Effectiveness of routine psychotherapy: Method matters. *Psychother Res* (2019) 2:139–56. doi: 10.1080/10503307.2017.1395921
14. Shadish WR, Matt GE, Navarro AM, Siegle G, Crits-Christoph P, Hazelrigg MD, et al. Evidence that therapy works in clincally representative conditions. *J Consult Clin Psychol* (1997) 65:355–65. doi: 10.1037/0022-006X.65.3.355
15. Katz AJ, de Krasinski M, Philip E, Wieser C. Change in interactions as a measure of effectiveness in short term family therapy. *Family Ther* (1975) 2:31–56.
16. Kazdin AE. Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behav Res Ther* (2017) 88:7–18. doi: 10.1016/j.brat.2016.06.004
17. Kirsch I, Huedo-Medina TB, Pigott HE, Johnson BT. Do outcomes of clinical trials resemble those "real world" patients? A reanalysis of the STAR* D antidepressant data set. *Psychol Consciousness: Theory Res Pract* (2018) 5:339–45. doi: 10.1037/cns0000164
18. Monsen K, Monsen JT. Chronic pain and psychodynamic body therapy: A controlled outcome study. *Psychother: Theory Res Practice Training* (2000) 37:257–69. doi: 10.1037/h0087658
19. Derogatis LR. *SCL-90. Administration, scoring and procedures. Manual for the R (revised) version and other instruments of the Psychopathology Rating Scale Series*. Baltimore: Johns Hopkins University School of Medicine (1992).
20. Horowitz LM, Rosenberg SE, Baer BA, Ureño G, Villaseñor VS. Inventory of interpersonal problems: Psychometric properties and clinical applications. *J Consult Clin Psychol* (1988) 56:885–92. doi: 10.1037/0022-006X.56.6.885
21. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders. 4th ed.* Washington DC: American Psychiatric Association (1994).
22. Derogatis LR, Savitz KL. The SCL-90-R, Brief Symptom Inventory, and Matching Clinical Rating Scales. In: Maruish M, editor. *The Use of*

Psychological Testing for Treatment Planning and Outcomes Assessment. Mahwah, N.J: Lawrence Erlbaum Associates Publishers. (2004) p. 679–724.
23. Hill CE, Lambert MJ. Methodological issues in studying psychotherapy processes and outcomes. In: Lamert M, editor. *Bergin and Garfield"s handbook of psychotherapy and behavior change, 6th ed*. New York, NY: John Wiley & Sons. (2013) p. 84–135.
24. Tracey TJG, Rounds J, Gurtman M. Examination of the General Factor with the Interpersonal Circumplex Structure: Application to the Inventory of Interpersonal Problems. *Multivariate Behav Res* (1996) 31:441–66. doi: 10.1207/s15327906mbr3104_3
25. Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clin Psychol Psychother* (2011) 18:75–9. doi: 10.1002/cpp.693
26. Core Team R. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (2014). Retrieved from http://www.R-project.org/.
27. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* (2015) 67:1–48. doi: 10.18637/jss.v067.i01
28. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw* (2017) 82:1–26. doi: 10.18637/jss.v082.i13
29. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, US: Springer-Verlag (2016). Available at https://CRAN.R-project.org/package=ggalluvial
30. Brunson JC. (2018). *ggalluvial: Alluvial Diagrams in "ggplot2". R package version 0.9.1*. Available at https://CRAN.R-project.org/package=ggalluvial
31. Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* (2010), 45:1–68. doi: 10.18637/jss.v045.i03
32. Hamer RM, Simpson PM. Last Observation Carried Forward Versus Mixed Models in the Analysis of Psychiatric Clinical Trials. *Am J Psychiatry* (2009) 166:639–41. doi: 10.1176/appi.ajp.2009.09040458
33. Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. *J Biopharmaceut Stat* (2001) 11:9–21. doi: 10.1081/BIP-100104194
34. Gallop R, Tasca GA. Multilevel modeling of longitudinal data for psychotherapy researchers: II. The complexities. *Psychother Res* (2009) 19:438–52. doi: 10.1080/10503300902849475
35. Goldberg SB, Hoyt WT, Nissen-Lie HA, Nielsen SL, Wampold BE. Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists. *Psychother Res* (2018) 28:532–44. doi: 10.1080/10503307.2016.1216625
36. Magnusson K, Andersson G, Carlbring P. The consequences of ignoring therapist effects in trials with longitudinal data: A simulation study. *J Consult Clin Psychol* (2018) 86:711–25. doi: 10.1037/ccp0000333
37. Wampold BE, Serlin RC. The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *psychol Methods* (2000) 5:425–33. doi: 10.1037/1082-989X.5.4.425
38. Sink CA, Mvududu NH. Statistical Power, Sampling, and Effect Sizes: Three Keys to Research Relevancy. *Couns Outcome Res Eval* (2010) 1:1–18. doi: 10.1177/2150137810373613
39. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum Associates (1988).
40. Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* (1991) 59:12–9. doi: 10.1037/0022-006X.59.1.12
41. Jacobson NS, Follette WC, Revenstorf D, Hahlweg K, Baucom DH, Margolin G. Variability in outcome and clinical significance of behavioral marital therapy: A reanalysis of outcome data. *J Consult Clin Psychol* (1984) 52:497–504. doi: 10.1037/0022-006X.52.4.497
42. Carrozzino D, Vassend O, Bjørndal F, Pignolo C, Olsen LR, Bech P. A clinimetric analysis of the Hopkins Symptom Checklist (SCL-90-R) in general population studies (Denmark, Norway, and Italy). *Nordic J Psychiatry* (2016) 70:374–9. doi: 10.3109/08039488.2016.1155235
43. Monsen JT, Hagtvet KA, Havik OE, Eilertsen DE. Circumplex structure and personality disorder correlates of the Interpersonal Problems Model (IIP-C): Construct validity and clinical implications. *psychol Assess* (2006) 18:165–73. doi: 10.1037/1040-3590.18.2.165
44. Fay MP. Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. *R J* (2010) 2:53–8. doi: 10.32614/RJ-2010-008
45. Fay MP. Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics* (2010) 11:373–4. doi: 10.1093/biostatistics/kxp050
46. Nordmo M. The Norwegian Multisite Study of Process and Outcome in Psychotherapy (NMSPOP) – Main Outcome Project. (2018). doi: 10.17605/OSF.IO/B5NWK
47. Silverman HJ, Miller FG. Control group selection in critical care randomized controlled trials evaluating interventional strategies: An ethical assessment*. *Crit Care Med* (2004) 32:852–7. doi: 10.1097/01.Ccm.0000114814.62759.06
48. Borkovec TD, Sibrava NJ. Problems with the use of placebo conditions in psychotherapy research, suggested alternatives, and some strategies for the pursuit of the placebo phenomenon. *J Clin Psychol* (2005) 61:805–18. doi: 10.1002/jclp.20127
49. Ghio L, Gotelli S, Marcenaro M, Amore M, Natta W. Duration of untreated illness and outcomes in unipolar depression: A systematic review and meta-analysis. *J Affect Disord* (2014) 52:45–51. doi: 10.1016/j.jad.2013.10.002
50. Merikangas K, Zhang H, Avenevoli S, Acharyya S, Neuenschwander M, Angst J. Longitudinal trajectories of depression and anxiety in a prospective community study: The zurich cohort study. *Arch Gen Psychiatry* (2003) 60:993–1000. doi: 10.1001/archpsyc.60.9.993
51. Bateman A, Fonagy P. 8-Year Follow-Up of Patients Treated for Borderline Personality Disorder: Mentalization-Based Treatment Versus Treatment as Usual. *Am J Psychiatry* (2008) 165:631–8. doi: 10.1176/appi.ajp.2007.07040636
52. Biskin RS. The Lifetime Course of Borderline Personality Disorder. *Can J Psychiatry* (2015) 60:303–8. doi: 10.1177/070674371506000702
53. Skodol AE, Gunderson JG, Shea MT, McGlashan TH, Morey LC, Sanislow CA, et al. The Collaborative Longitudinal Personality Disorders Study (CLPS): Overview and Implications. *J Pers Disord* (2005) 19:487–504. doi: 10.1521/pedi.2005.19.5.487
54. Zanarini MC, Frankenburg FR, Vujanovic AA, Hennen J, Reich DB, Silk KR. Axis II comorbidity of borderline personality disorder: description of 6-year course and prediction to time-to-remission. *Acta Psychiatr Scand* (2004) 110:416–20. doi: 10.1111/j.1600-0447.2004.00362.x
55. Bender DS, Dolan RT, Skodol AE, Sanislow CA, Dyck IR, McGlashan TH, et al. Treatment Utilization by Patients With Personality Disorders. *Am J Psychiatry* (2001) 158:295–302. doi: 10.1176/appi.ajp.158.2.295
56. Shean G. Psychotherapy Outcome Research: Issues and Questions. *Psychodynamic Psychiatry* (2016) 44:1–24. doi: 10.1521/pdps.2016.44.1.1
57. Finnes A, Enebrink P, Ghaderi A, Dahl J, Nager A, Öst LG, et al. Psychological treatments for return to work in individuals on sickness absence due to common mental disorders or musculoskeletal disorders: a systematic review and meta-analysis of randomized-controlled trials. *Int Arch Occup Environ Health* (2018), 92:273–93. doi: 10.1007/s00420-018-1380-x
58. Henderson M, Harvey S, Øverland S, Mykletun A, Hotopf M. Work and common psychiatric disorders. *J R Soc Med* (2011) 104:198–207. doi: 10.1258/jrsm.2011.100231
59. Løvvik C, Shaw W, Øverland S, Reme SE. Expectations and illness perceptions as predictors of benefit recipiency among workers with common mental disorders: secondary analysis from a randomised controlled trial. *BMJ Open* (2014) 4:3. doi: 10.1136/bmjopen-2013-004321
60. Wasiak R, Young AE, Roessler RT, McPherson KM, van Poppel MNM, Anema JR. Measuring Return to Work. *J Occup Rehabil* (2007) 17:766–81. doi: 10.1007/s10926-007-9101-4
61. Cuijpers P, Li J, Hofmann SG, Andersson G. Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clin Psychol Rev* (2010) 30:768–78. doi: 10.1016/j.cpr.2010.06.001
62. Ogles BM. Measuring change in psychotherapy research. In: Lambert M, editor. *Bergin & Garfield"s handbook of psychotherapy and behavior change, 6th ed.* New York, NY: John Wiley & Sons. (2013) p. 134–66.
63. Hansen NB, Lambert MJ, Forman EM. The Psychotherapy Dose-Response Effect and Its Implications for Treatment Delivery Services. *Clin Psychol: Sci Pract* (2002) 9:329–43. doi: 10.1093/clipsy.9.3.329
64. Vasquez MJT, Bingham RP, Barnett JE. Psychotherapy termination: clinical and ethical responsibilities. *J Clin Psychol* (2008) 64:653–65. doi: 10.1002/jclp.20478

65. Alsawy S, Mansell W, Carey TA, McEvoy P, Tai SJ. Science and practice of transdiagnostic CBT: a Perceptual Control Theory (PCT) approach. *Int J Cogn Ther* (2014) 7:334–59. doi: 10.1521/ijct.2014.7.4.334
66. Baldwin SA, Imel ZE. Therapist effects: Findings and methods. In: Lambert M, editor. *Bergin and Garfield"s handbook of psychotherapy and behavior change, 6th ed.* New York, NY: John Wiley & Sons. (2013). p. 258–97.
67. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *Int Rev Psychiatry* (2011) 23:318–27. doi: 10.3109/09540261.2011.606803
68. McHugh RK, Barlow DH. The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *Am Psychol* (2010) 65:73–84. doi: 10.1037/a0018121
69. Knapstad M, Nordgreen T, Smith OR. Prompt mental health care, the Norwegian version of IAPT: clinical outcomes and predictors of change in a multicenter cohort study. *BMC Psychiatry* (2018) 18:260. doi: 10.1186/s12888-018-1838-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

II

III

Appendix

Paper 1 Registration

Research Questions

In recent years we have witnessed a substantial development and expansion of the knowledge base in psychotherapy research. The fundamental effectiveness of psychological treatment, the dose-responsiveness relationship, the differential effectiveness of therapists, the equivalence of outcomes across systematic treatment models, the supremacy of focused as opposed to purely supportive interventions, and the superiority of long-term, high intensity treatment for patients with severe characterological disorder have all been firmly established. Still, the majority of studies demonstrating these findings have been on patients with, relatively speaking, lower levels of disorder complexity and psychological dysfunction, often excluding those with multiple comorbid disorders and severe character-based pathology, in relatively short-term treatment. Thus, there is a notable absence of studies demonstrating the effectiveness of common treatment approaches with more complex patient-populations. Since the majority of patients receiving specialist psychotherapy services can be classified as complex cases with multiple disorders and commonly have personality-based pathology in addition to their symptom disorders, this is a very unfortunate limitation in the existing literature.

Our dataset (The Norwegian Multisite Study of Process and Outcome in Psychotherapy – NMSPOP) is designed to remedy these limitations. It includes naturalistically selected patients from clinical specialist services, half of which satisfy criteria for at least one personality disorder at treatment start. The mean number of Axis I disorders at treatment start is 2.01, demonstrating the commonality of case complexity. Furthermore, the NMSPOP is designed so that the majority of treatments delivered are in the open-ended format of psychotherapy (i.e. therapist and patient are instructed to come to an agreement about when to terminate treatment based on the patients difficulties and progress or lack thereof). Our goal is to demonstrate the effects of psychotherapy within this context on both self-rated questionnaires, completed by the participants, and also in terms of observer-rated diagnostic changes. We also want to examine changes in occupational activity as patient's progress through their therapy. Previous research has outlined that psychosocial interventions that do not specifically target occupational status, can be successful in treating therapeutic outcomes, and still fail to produce significant changes in occupational status.

Hypotheses

For each of the research questions listed in the previous section, provide one or multiple specific and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here.

Symptoms

1) We predict a large reduction (Cohen's d > .8) on the patient-rated Symptom Checklist-90 (SCL- 90): Global Severity Index (GSI) from pre-treatment to treatment completion.
2) We predict that reductions achieved on the GSI will be maintained throughout the follow-up period, with the last measure taken at 2.5 years after treatment completion.
3) We predict a large reduction (Cohen's d > .8) in observer-rated symptom disorders (as assessed with the SCID-I) from pre-treatment to treatment completion.
4) We predict that the reductions in observer-rated symptom disorders will be maintained throughout the follow-up period, with the last assessment conducted 2.5 years after treatment completion.
5) We predict that a majority of patients will be classified as either "recovered" or "improved" based on a clinical significance classification of the GSI from pre-treatment to treatment completion
6) We predict that a majority of patients will be classified as either "recovered" or "improved" based on a clinical significance classification of the GSI from pre-treatment to 2.5 years follow-up.

Interpersonal

7) We predict a moderate reduction (Cohen's d > .5) on the Inventory of Interpersonal Problems 64 (IIP-64) global score from pre-treatment to treatment completion.
8) We predict that reductions achieved on the IIP-64 global score will be maintained throughout the follow-up period, with the last measure conducted 2.5 years after treatment completion
9) We predict a large reduction (Cohen's d > .8) in observer-rated personality disorder diagnoses (as assessed with the SCID-II) from pre-treatment to treatment completion.
10) We predict that observer-rated reductions in personality disorders will be maintained through the follow-up period, with the last assessment conducted 2.5 years after treatment completion.
11) We predict that a majority of patients will be classified as either "recovered" or "improved" based on a clinical significance classification of the IIP Global from pre-treatment to treatment completion
12) We predict that a majority of patients will be classified as either "recovered" or "improved" based on a clinical significance classification of the IIP Global from pre-treatment to 2.5 year follow-up

Occupational functioning

13) We predict that there will be a moderate but statistically significant increase in work functioning, with higher functioning after treatment completion, compared to pre-treatment levels.
14) We predict that the positive gains in work functioning will be maintained through the follow-up period, with the last measure taken 2.5 years after treatment completion.

Explanation of Existing Data

One of the main goals of this project was to assess naturalistic psychotherapy outcomes from a sample with a flexible framework and a heterogeneous patient population. That was a goal when the project was started, and it`s a goal now. The original project was started in 1995, and data gathering was concluded in 2008. A number of scientific process-oriented publications have been published with results from the dataset, but no quantitative analysis of the main outcomes has been analyzed or published. There are a number of reasons for this delay. Most of them pertain to organizational obstacles, as well as challenges with assessing inter-rater reliability on diagnoses.

See the link for a list of publications. http://www.sv.uio.no/psi/forskning/prosjekter/multisenter/publikasjoner/index.html. Some of the authors have made informal analyses of the results prior to preregistration. These results have been presented at the 39th International Meeting of the Society for Psychotherapy Research in Barcelona, June 18-21, 2008. Under the following title: "The Norwegian Multisite Study of Process and Outcome in Psychotherapy – Some preliminary analyses of outcome." Every relevant outcome variable was presented but with a less sophisticated methodology than we are currently pursuing. We do not believe that these preliminary analyses will affect our planned analyses. Effectiveness research is by nature confirmatory as the null hypothesis is always that the treatment is not effective. We have made deliberate attempts to secure that previous informal analyses do not influence our current hypotheses nor our choice of statistical procedures. The first author have not assessed the data. The goal of this preregistration is to provide transparency into the statistical analyses

of our data. We acknowledge that this late registration is not optimal, but maintain that the added transparency makes the preregistration valuable nonetheless.

Data Collection Procedures

The total NMSPOP sample consists of 370 adult outpatients from nine psychiatric treatment sites. The patients were selected based on eligibility from a pool of regularly referred outpatients, and participation was based on informed and signed consent. Participant enrollment lasted from May 1995, throughout December 2000. Treatment was completed for every patient between the period of March 1996 to March 2007. At each of the nine sites, a specially trained coordinator (clinical psychologist or psychiatrist) was responsible for recruiting patients. The coordinators were instructed to select randomly from their local patient population but to ensure that roughly half had a personality disorder. The same coordinators carried out the assessment of the patients. Patients with serious substance abuse, psychoses, and acute crises requiring hospitalization as well as those with an IQ below 70 and who were younger than 20 years were not included in the study. The patients represented a heterogeneous and typical outpatient sample. Patients were not compensated for participating in the study. Every patient will be included in our analyses. N = 370

Measured Variables

Self-rated symptom scales:

- Symptom Checklist 90: General Severity Index (GSI)
- Inventory of Interpersonal Problems 64: IIP Global

Clinician-rated:

- SCID Axis I: Symptom Disorder
- SCID Axis II: Personality Disorder

Other: Self-reported occupational status

Single item question: "Describe your current occupational-status" with seven possible alternatives

1. I am currently engaged in paid work
2. I am a "stay at home" mom/dad
3. I am currently engaged as a student
4. I am currently on sick-leave
5. I am currently in work rehabilitation
6. I am currently receiving disability benefit
7. I am currently unemployed

Indices

Clinical Significance. We will use Jacobsen and Truax (1991) definition of clinical significance to assess therapeutic gains. Clinically significant change occurs when a patient moves from a dysfunctional population to a functional or normal population during treatment and the magnitude of that patient's change is statistically reliable. A patient whose improvement meets both of these criteria is classified as recovered (having returned to normal functioning). On the basis of these criteria, patients are categorized as (1) recovered, (2) reliably improved but not recovered, (3) unchanged, or (4) deteriorated, in the case of reliable negative change.

Effect size

Effect Size. We will use Cohens (1988) standards for evaluating the magnitude of effect sizes, classifying small effects as d = 0.2–0.5; medium effects as d =0.5–0.8; and large effects as d = > .8. In order not to underestimate error and inflate effect sizes, estimated changes will be divided by the pooled standard deviations of all relevant measurement points on the outcome variables. Thus, the pooled standard deviations of estimated scores across all measurement points on each outcome variable will be used when estimating the effect sizes.

Work status. Self-reported work status will be indexed into a generic "functioning" variable based on the following responses:

"Functioning":

- I am currently engaged in paid work
- I am a "stay at home" mom/dad
- I am currently engaged as a student

"Non-functioning":

- I am currently on sick-leave
- I am currently in work rehabilitation
- I am currently receiving disability benefit
- I am currently unemployed

Diagnostic status. Diagnoses are assessed by a clinician using Structured Clinical Interview for DSM 4, AXIS I & II. The following indexes will be used when reporting diagnoses:

Axis 1:

0: No symptomatic diagnosis during the last month

1: One or more symptomatic diagnoses during the last month

Axis 2:

0: No personality disorder diagnosis during the last month

1: One or more personality disorder diagnoses during the last month

SCL 90-R: Global severity index. Global severity index is the mean score across every item in the SCL-90.

IIP-64: Global score. The global score is the mean score across every item in the IIP-64.

Statistical Models

Symptom measures (SCL-90: GSI and IIP-64 Global) will be analyzed separately using linear mixed-effect models to estimate patient trajectories, using Time as a predictor. Repeated measures (level 1) will be nested within individuals (level 2) which allows the model to assume non-independence with repeated data. Linear mixed-effect models also allow for the inclusion of all available data, making this an intent-to-treat analysis. Each of the two outcome variables will be divided into two separate models. One for assessing therapeutic gains from pretreatment, to during- and post-treatment. Another to assess whether treatment gains are maintained from post-treatment, to 6 months follow-up (FU), to 1 year FU, to 2.5 years FU. To maximize fit we will use an unstructured covariance structure.

Diagnosis status (Axis I and II) will be assessed using separate McNemar $\chi 2$ -tests for longitudinal analysis of related dichotomous variables. The presence of a diagnosis on either axis is recorded before treatment onset, after treatment completion, 6 months after treatment completion, 1 year after treatment completion and 2.5 years after treatment completion.

Symptom measures (SCL 90-R and IIP 64) will be centered at time level zero in order to interpret intercept values. Model fit will be assessed empirically using Bayesian Information Criterion (BIC).

Data exclusion

A goal of this publication is to publish the total results from the entire NMSPOP project. Thusly, every participants and relevant data point will be included in our intention-to-treat analysis with a single exception: As a part of another research project, an extra (n = 23)

sample was collected roughly four years after treatment termination. This extra sample will be disregarded. Linear mixed-effect models are equipped to handle missing data.

Paper 2 Registration

Research Questions

There is a substantial literature supporting the notion that psychotherapy produces meaningful positive change for individuals with a mental illness (Dragioti, Karathanos, Gerdle, & Evangelou, 2017). The majority of this research literature is, however, based on samples with limited psychopathology, often excluding those with multiple comorbid disorders and severe character based pathology (Cooper & Conklin, 2015; Hoertel et al., 2014; Hoertel et al., 2015; Lambert, 2013; Ronconi, Shiner, & Watts, 2014; Shadish, Navarro, Matt, & Phillips, 2000). More often than not, this research examine a relatively short-term and structured treatment protocol focused on a specific DSM-defined symptom (Shean, 2014). The literature that specially cover treatment interventions for patient with a personality disorder mostly focuses on Borderline Personality Disorder (BPD) and is generally limited by small sample sizes, short follow-up periods and poor control of coexisting psychopathology (Bateman, Gunderson, & Mulder, 2015). Our dataset (The Norwegian Multisite Study of Process and Outcome in Psychotherapy – NMSPOP) is designed to remedy these limitations. It includes naturalistically selected patients from routine outpatient care, half of which satisfy DSM-4 (American Psychiatric Association, 1994) criteria for at least one personality disorder at treatment start. The treatments provided in the NMSPOP trial are open-ended, i.e. therapist and patient are instructed to come to an agreement about when to terminate treatment. Our goal is to investigate the potentially moderating effect of personality pathology on overall treatment effect of psychotherapy. This will be investigated in terms of both self-rated questionnaires, completed by the participants, and also with observer-rated diagnostic changes. This registration is the second in the NMSPOP project. The first focused exclusively on total sample treatment effects. The first round of analysis (not published as of this writing) has already verified that our sample, taken as a whole, experiences large and moderate changes in psychiatric symptoms and interpersonal problems, respectively. We also asses individual patient data using clinical significant change (Jacobson & Truax, 1991) and find that, following treatment, a sizeable majority either recovered (38 %) or improved (31 %) when assessing symptomatic change. When assessing interpersonal problems, we found that 23 % recovered and 12 % improved. Most patients maintained these changes at the two-and-a-half-year follow-up.

Hypotheses

Given the lack of research in this area, we are unsure of the moderating effect of personality disorder on the effect of open-ended psychotherapy. However, one of the goals of the NMSPOP was to asses a clinical intuition, namely that patients with severe personality pathology can benefit from psychotherapy if they are given flexible treatment conditions that allow for individualized treatment protocols designed by each respective clinician in collaboration with the patient. Therefor we predict that the presence of a personality disorder will not moderate the changes seen in patients undergoing psychotherapy.

Explanation of Existing Data

The original project was started in 1995, and data gathering was concluded in 2008. A number of scientific process-oriented publications have been published with results from the dataset. See the link for a list of publications. http://www.sv.uio.no/psi/forskning/prosjekter/multisenter/publikasjoner/index.html. Some of the authors have made informal analyses of the results prior to this registration. These results have been presented at the 39th International Meeting of the Society for Psychotherapy Research in Barcelona, June 18-21, 2008. Under the following title: "The Norwegian Multisite Study of Process and Outcome in Psychotherapy – Some preliminary analyses of outcome." This analysis was limited to main outcomes and did not include any moderator analysis. However, some of the authors of this current paper has made informal descriptive (means, data visualization) investigations of personality pathology as predictor of change. We have made deliberate attempts to secure that previous investigations do not influence our current hypotheses nor our choice of statistical procedures. The goal of this registration is to provide transparency into the statistical analyses of our data. Therefore, we feel the term "pre-registration" is misleading and apply a more conservative "registration", to indicate that we have accessed the data before writing completing this registration.

Data Collection Procedures

The total NMSPOP sample consists of 370 adult outpatients from eight psychiatric treatment sites. The patients were selected based on eligibility from a pool of regularly referred outpatients, and participation was based on informed and signed consent. We only used data from routine outpatient facilities in our analysis. We collected the majority of patient (n = 301) data from psychiatric outpatient clinics spread across 17 separate clinics. The clinics are nested in six over-arching treatment sites, which correspond to different hospital regions in Norway. We also gathered data from the Norwegian University of Science and Technology`s student clinic (n = 27). Lastly, we gathered data from an outpatient clinic (n = 42) specializing in psychodynamic body therapy for patient with somatoform disorders. Participant enrollment lasted from May 1995, throughout December 2000. Treatment was completed for every patient between the period of March 1996 to March 2007. At each of the eight sites, a specially trained coordinator (clinical psychologist or psychiatrist) was responsible for recruiting patients. The coordinators were instructed to select randomly from their local patient population but to ensure that roughly half had a personality disorder. The same coordinators carried out the assessment of the patients. Patients with serious substance abuse, psychoses, and acute crises requiring hospitalization as well as those with an IQ below 70 and who were younger than 20 years were not included in the study. The patients represented a heterogeneous and typical outpatient sample. Patients were not compensated for participating in the study. We plan to include every patients in our analyses: N = 370.

Measured Variables

Self-rated symptom scales. To assess symptomatic change, we will utilize the Symptom Checklist 90: General Severity Index (GSI). The GSI is the mean score across every item in the SCL-90. To assess changes in interpersonal functioning we will utilize the Inventory of Interpersonal Problems 64: IIP Global. The IIP Global is the mean score across every item in the IIP 64.

Observer-rated diagnoses. Diagnoses are assessed by an observer using Structured Clinical Interview for DSM 4 (American Psychiatric Association, 1994). We will apply a dichotomous coding scheme for diagnostic status.

Structured Clinical Interview for DSM 4 Axis I: Clinical Disorders (SCID I). The SCID interview will be assessed as a dichotomous diagnose (1) or no diagnose (0) measure. A score of 1 means that the patient has qualified for one or more clinical disorders during the last month. A score of 0 indicates that the patient has not qualified for any of the DSM 4 clinical disorders during the last month.

Structured Clinical Interview for DSM 4 Axis II: Personality Disorder (SCID II). The SCID interview will be assessed as a dichotomous diagnose (1) or no diagnose (0) measure. A score of 1 means that the patient has qualified for one or more personality disorders during the last month. A score of 0 indicates that the patient has not qualified for any of the DSM 4 personality disorders during the last month.

Indices

Clinical Significance. We will use Jacobsen and Truax (1984) definition of clinical significance to assess therapeutic gains. Clinically significant change occurs when a patient moves from a dysfunctional population to a functional or normal population during treatment and the magnitude of that patient's change is statistically reliable. A patient whose improvement meets both of these criteria is classified as recovered (having returned to normal functioning). On the basis of these criteria, patients are categorized as (1) recovered, (2)

reliably improved but not recovered, (3) unchanged, or (4) deteriorated, in the case of reliable negative change. We plan on using predicted values from our multilevel regression models to assess clinical significance.

Effect size. We will use Cohens (1988) standards for evaluating the magnitude of effect sizes, classifying small effects as d = 0.2–0.5; medium effects as d = 0.5–0.8; and large effects as d > .8. In order not to underestimate error and inflate effect sizes, estimated changes will be divided by the pooled standard deviations of all measurement points with more than 150 participants for each respective outcome variable. We plan on using predicted values to estimate standard deviation and compare means.

Main Self-report Outcome Analyses

The dependent variables will be SCL-90: GSI and IIP-64 Global. Each will be analyzed separately using linear mixed-effect models to estimate patient trajectories, using Time as a predictor and a personality disorder related measures as moderators. We will apply session number as a fixed occasions time-estimate as this allows for between subject comparisons of regression coefficients. Post measurements for each patient will be coded by assigning the last session number from the longest treatment series, plus one, which is 361. We will code time as zero at treatment completion and used month after treatment completion as our time measure for the follow-up phase. Repeated measures (level 1) will be nested within individuals (level 2) which allows each model to assume non-independence with repeated data. Linear mixed-effect models also allow for the inclusion of all available data, making this an intent-to-treat analysis. To maximize fit we plan to use an unstructured covariance. Model fit will be assessed empirically using Bayesian Information Criterion (BIC). Each outcome variable will be further divided into a treatment phase and a follow-up phase. The main research question is: Does the presence of a SCID II personality diagnosis pretreatment moderate the changes seen in the treatment and follow-up phase?

1. Self-report measures Pre- to Posttreatment, including all during treatment measures, as the Dependent Variable, with Time and Personality Disorder Status interaction as the Independent Variable. Question: Does the presence of a SCID II personality diagnosis pretreatment moderate the changes seen in the follow-up phase?
2. Self-report measures Posttreatment to last Follow-Up assessment as the Dependent Variable, with Time and Personality Disorder Status interaction as the Independent Variable.

Missing self-report data. Linear mixed-effect models are equipped to handle missing data as the models will include every relevant data when assessing individuals as a separate measurement level (Tasca & Gallop, 2009). We do not plan to use any other tools to handle missing data.

Diagnostic changes. Does the presence of a SCID II personality diagnosis moderate the changes seen in SCID I Clinical Disorders during the treatment phase.We plan to apply a Chi-Square test to assess whether presence of a SCID II Personality Diagnosis pretreatment is associated with the level of change seen in SCID I Clinical Diagnosis (CD). Only patients with a SCID I CD pretreatment will be included in this analysis. We will apply the following dummy coding scheme to construct our CD change variable:

0) The patient lost his/her SCID CD
1) The patients SCID I CD remained

Two separate analysis will be performed to compare pretreatment and posttreatment, as well as pretreatment to follow-up.

Missing Diagnostic Data. We plan to handle missing diagnostic data with two separate methods. Firstly, we will use listwise deletion which exclude patients that have missing diagnostic. Secondly, we plan on using the Multivariate Imputation by Chained Equations (MICE) methodology to impute missing diagnostic data (Buuren & Groothuis-Oudshoorn, 2010).

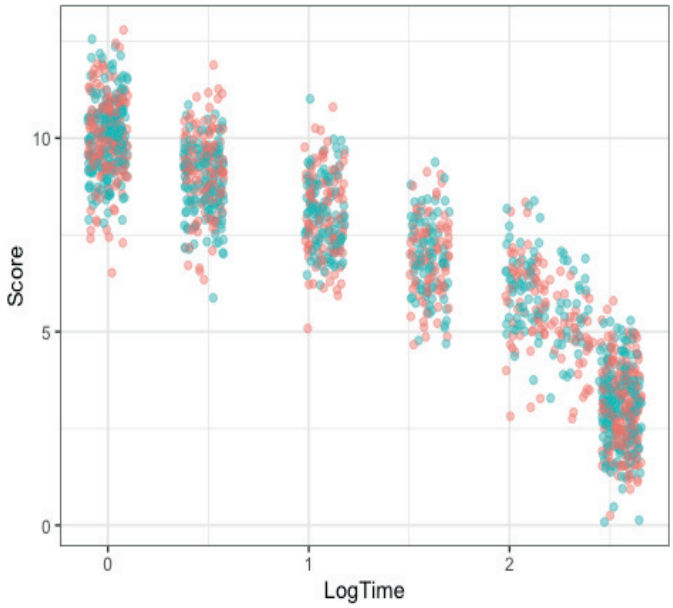
Planned Analyses using Simulated Data

```
library(tidyverse)
library(broom)
library(lme4)
library(broom.mixed)
library(lmerTest)
library(ggeffects)
library(MASS)


#####Simulate data with dependency
n <- 370
r <- .09
data <- mvrnorm(n, mu<-c(10,9,8,7,6,5,4,3),
               Sigma<-matrix(c(1,rep(r,8),1,rep(r,8), 1,rep(r,8),1,rep(r,8),1,
                              rep(r,8),1,rep(r,8),1,rep(r,8),1),
                             nrow=8,ncol = 8), empirical=TRUE)
###########
id <- c(1:n)
"Time0" <- data[, 1]
"Time1" <- data[, 2]
"Time2" <- data[, 3]
"Time3" <- data[, 4]
"Time4" <- data[, 5]
"Time5" <- data[, 6]
"Time6" <- data[, 7]
"Time7" <- data[, 8]
"PD" <- sample(0:1,n, replace = T)
"CD" <- sample(0:1,n, replace = T)
###########

df <- data.frame(id,Time0,Time1,Time2,Time3,Time4,Time5,
                 Time6,Time7,PD,CD)
rm(id,data,Sigma,Time0,Time1,Time2,Time3,Time4,Time5,
   Time6,Time7,PD,CD,r,mu,n)
###### Introduce NA to patients with short treatments (based on id)
```

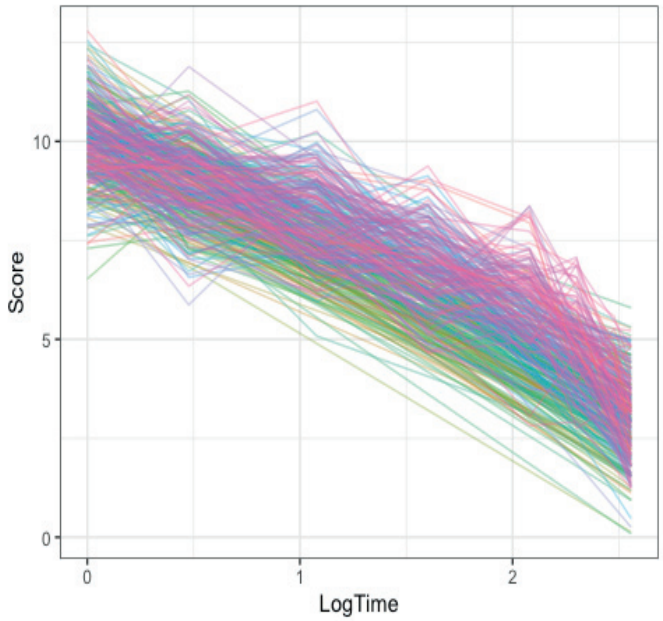
```r
x <- df %>%
  dplyr::select(id,Time0,Time7,PD,CD)
a <- df %>%
  dplyr::select(id,Time1) %>%
  filter(id > 100)
x <- left_join(x,a, key = "id")
a <- df %>%
  dplyr::select(id,Time2) %>%
  filter(id > 150)
x <- left_join(x,a, key = "id")
a <- df %>%
  dplyr::select(id,Time3) %>%
  filter(id > 200)
x <- left_join(x,a, key = "id")
a <- df %>%
  dplyr::select(id,Time4) %>%
  filter(id > 250)
x <- left_join(x,a, key = "id")
a <- df %>%
  dplyr::select(id,Time5) %>%
  filter(id > 300)
x <- left_join(x,a, key = "id")
a <- df %>%
  dplyr::select(id,Time6) %>%
  filter(id > 350)
x <- left_join(x,a, key = "id")
####################### Make it tidy
df <- x[,c(1,2,6:11,3,4,5)]
df <- df %>% gather("Time","Score",2:9)
df$Time <- str_replace(df$Time,"Time","")
df$Time <- as.integer(df$Time)
df <- df[order(df$id,df$Time),]
numbers <- c(0.1,0.3,0.5,0.7,0.9,1.1,1.3)
df$Score[df$Score < 0] <- sample(numbers,1)
df <- df %>%
  filter(!is.na(Score))
rm(x,a,numbers)
###################
df$Time <- recode(df$Time,"0" = 0, "1" = 3, "2" = 12,"3" = 40, "4" = 120, "5" = 200,
                  "6" = 300, "7" = 361)
df$LogTime <- ifelse(df$Time > 0, log10(df$Time),0)
df$PD <- as.factor(df$PD)
df$CD <- as.factor(df$CD)
summary(df)
##        id         PD      CD           Time           Score
##  Min.   :  1.0   0:797   0:775   Min.   :  0.0   Min.   : 0.0876
##  1st Qu.:164.0   1:813   1:835   1st Qu.:  3.0   1st Qu.: 4.7683
##  Median :251.0                   Median : 12.0   Median : 7.5878
##  Mean   :232.7                   Mean   :110.7   Mean   : 7.0431
##  3rd Qu.:316.0                   3rd Qu.:200.0   3rd Qu.: 9.3391
##  Max.   :370.0                   Max.   :361.0   Max.   :12.7911
##     LogTime
##  Min.   :0.0000
##  1st Qu.:0.4771
##  Median :1.0792
##  Mean   :1.2702
##  3rd Qu.:2.3010
##  Max.   :2.5575
```

###Visual Inspection

```r
ggplot(df,aes(x = LogTime, y = Score)) +
  geom_jitter(aes(color = PD),alpha = 0.4,width = 0.1) +
  theme_bw() +
  theme(legend.position="none")
```



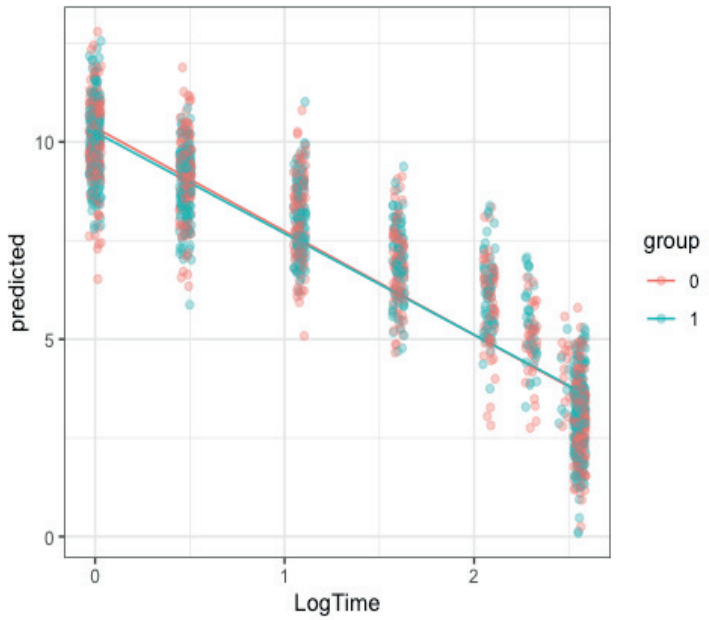

```
ggplot(df,aes(x = LogTime, y = Score, group = id)) +
  geom_line(aes(color = as.factor(id)),alpha = 0.3) +
  theme_bw() +
  theme(legend.position="none")
```




```
###Statistical Analyses of Self-Report Outcomes
```

```
mlmodell <- lmer(Score ~ LogTime * PD + (LogTime | id), data=df)
## boundary (singular) fit: see ?isSingular
## Warning: Model failed to converge with 1 negative eigenvalue: -3.0e+01
summary(mlmodell)
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Score ~ LogTime * PD + (LogTime | id)
##    Data: df
##
## REML criterion at convergence: 5022.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4595 -0.6524 -0.0166  0.6566  3.0500
##
```

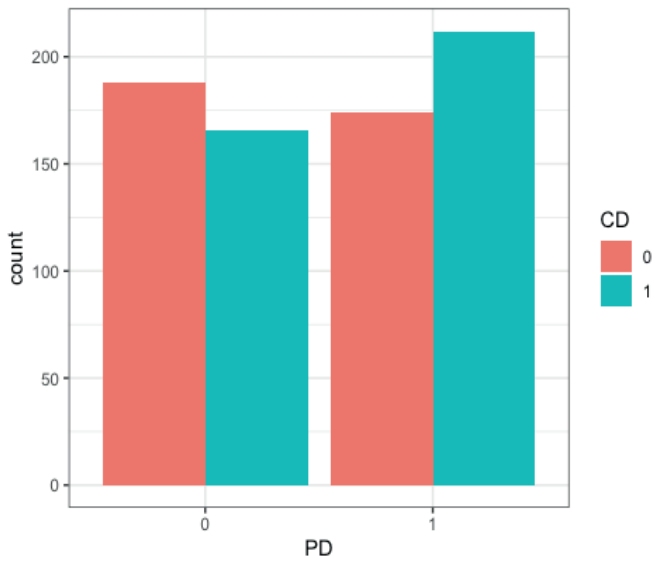
```
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.00000  0.0000
##           LogTime     0.03646  0.1909    NaN
##  Residual             1.23208  1.1100
## Number of obs: 1610, groups:  id, 370
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)   10.36343    0.06448 1277.05886 160.727   <2e-16 ***
## LogTime       -2.62510    0.04277  579.12124 -61.372   <2e-16 ***
## PD1           -0.11233    0.08999 1280.30679  -1.248    0.212
## LogTime:PD1    0.05547    0.05966  592.72932   0.930    0.353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) LogTim PD1
## LogTime     -0.736
## PD1         -0.717  0.527
## LogTime:PD1  0.528 -0.717 -0.733
## convergence code: 0
## boundary (singular) fit: see ?isSingular
mydf <- ggpredict(mlmodell,terms = c("LogTime","PD"))

ggplot(mydf, aes(x, predicted, color = group)) + geom_line() +
  geom_jitter(aes(y = Score,x = LogTime,color = PD),alpha = 0.3, data = df) +
  xlab("LogTime") +
  theme_bw()
```



Statistical Analyses of Diagnostic Outcomes

```
# Only include Pre and Post
PrePost <- df %>%
  filter(Time %in% c(0,361)) %>%
  dplyr::select(c(PD,CD))

Table <- table(PrePost$PD,PrePost$CD)

colnames(Table) <- c("No Personality Disorder", "Personality Disorder")

rownames(Table) <- c("SCID I Diagnosis Lost", "SCID I Diagnosis Remained")

Table
##
##                              No Personality Disorder Personality Disorder
##   SCID I Diagnosis Lost                         188                  166
##   SCID I Diagnosis Remained                     174                  212
```

```
chisq.test(Table)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Table
## X-squared = 4.4485, df = 1, p-value = 0.03493
# Visual inspection
ggplot(df %>% filter(Time %in% c(0,361)), aes(PD, ..count..)) +
  geom_bar(aes(fill = CD), position = "dodge") +
  theme_bw()
```



References

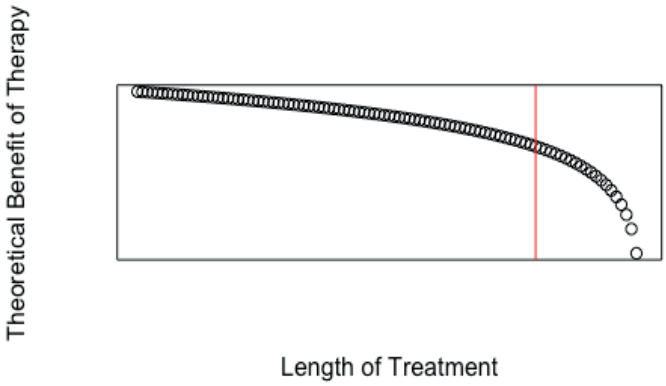
- American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author
- Bateman, A. W., Gunderson, J., & Mulder, R. (2015). Treatment of personality disorder. The Lancet, 385(9969), 735-743. https://doi.org/10.1016/S0140-6736(14)61394-5
- Buuren, S. v., & Groothuis-Oudshoorn, K. J. (2010). mice: Multivariate imputation by chained equations in R. 1-68.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, N.J.: L. Erlbaum Associates.
- Cooper, A. A., & Conklin, L. R. (2015). Dropout from individual psychotherapy for major depression: A meta-analysis of randomized clinical trials. Clinical Psychology Review, 40, 57-65. https://doi.org/10.1016/j.cpr.2015.05.001
- Dragioti, E., Karathanos, V., Gerdle, B., & Evangelou, E. (2017). Does psychotherapy work? An umbrella review of meta-analyses of randomized controlled trials. 136(3), 236-246. https://doi.org/10.1111/acps.12713
- Hoertel, N., de Maricourt, P., Katz, J., Doukhan, R., Lavaud, P., Peyre, H., & Limosin, F. (2014). Are Participants in Pharmacological and Psychotherapy Treatment Trials for Social Anxiety Disorder Representative of Patients in Real-Life Settings?, 34(6), 697-703. https://doi.org/10.1097/jcp.0000000000000204
- Hoertel, N., López, S., Wang, S., González-Pinto, A., Limosin, F., & Blanco, C. (2015). Generalizability of pharmacological and psychotherapy clinical trial results for borderline personality disorder to community samples. Personality disorders, 6(1), 81-87. https://doi.org/10.1037/per0000091
- Jacobson, N. S., Follette, W. C., Revenstorf, D., Hahlweg, K., Baucom, D. H., & Margolin, G. (1984). Variability in outcome and clinical significance of behavioral marital therapy: A reanalysis of outcome data. Journal of Consulting and Clinical Psychology, 52(4), 497-504. https://doi.org/10.1037/0022-006X.52.4.497
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59(1), 12-19. https://doi.org/10.1037/0022-006X.59.1.12
- Lambert, M. J. (2013). The efficacy and effectivness of Psychotherapy. In M. J. Lambert (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, 6th edition. (pp. 93-139). N.J. Hoboken: John Wiley & Sons.
- Ronconi, J. M., Shiner, B., & Watts, B. V. (2014). Inclusion and exclusion criteria in randomized controlled trials of psychotherapy for PTSD. J Psychiatr Pract, 20(1), 25-37. https://doi.org/10.1097/01.pra.0000442936.23457.5b
- Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. Psychological Bulletin, 126(4), 512-529. https://doi.org/10.1037/0033-2909.126.4.512
- Shean, G. (2014). Limitations of Randomized Control Designs in Psychotherapy Research %J Advances in Psychiatry. 2014, 5. https://doi.org/10.1155/2014/561452
- Tasca, G. A., & Gallop, R. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. Psychotherapy Research, 19(4-5), 429-437. https://doi.org/10.1080/10503300802641444

Paper 3 Registration

Research Question

The question of how much therapy is needed for psychotherapy to produce its effects has been hotly debate since the inception of psychotherapy itself. The first empirical investigators likened dosage of psychotherapy to a pharmaceutical drug and found a log-linear dose-effect relationship (Howard et al., 1986). Data from psychotherapy trials revealed that the majority of improvements were seen at early phase

of therapy, usually within the first 8 sessions. If a patient failed to achieve early improvement, then the likelihood of establishing it at a later phase decreased at a log-linear rate. An “optimal dose”, i this view, is a theoretical point of rapidly diminishing returns.



The general pattern of log-linear improvement has been replicated across several RCTs and in naturalistic settings (Robinson et al., 2019). Different settings offer heterogeneous data sources with differences related to patients, settings and treatments. For example, data from University Counseling Centers suggest that the optimal dose could be between 4-11 sessions (Anderson & Lambert, 2001, Draper et al., 2002, Robinson et al., 2019). Another perspective was provided by Baldwin (2009) who replaced the pharmaceutical analogy with a focus on the psychotherapeutic process. The *good-enough level (GEL)* model suggest that patients differ in their response to psychotherapy, with some patients rapidly improving, while others needing more therapy to achieve the same results. The GEL model posits that each patients attends psychotherapy for as long as is required to achieve a “good enough” level of functioning. Once this level is achieved the patient or therapist terminate the treatment. This model can also explain the previously established log-linear pattern of improvement as the majority of patients improve quickly and then terminate treatment. Our data set (The Norwegian Multisite Study of Process and Outcome in Psychotherapy – NMSPOP) is appropriate to assess and contrast the dose-response and the GEL model. Unlike the majority of previous studies, it includes naturalistically selected patients from routine outpatient care. Patients are, on average, more severely inflicted with psychopathology, as half of of the patients satisfy DSM-4 (American Psychiatric Association, 1994) criteria for at least one personality disorder at treatment start. The patients also have a high degree of co-morbidity and the majority has been afflicted with psychopathology for several years. Lastly, in contrast to previous studies, treatments provided in the NMSPOP trial are open-ended, i.e. therapist and patient are instructed to come to an agreement about when to terminate treatment.

This registration is the third in a set of NMSPOP project. The first, (https://osf.io/aq4vg/) focused exclusively on total sample treatment effects. These analyses (not published as of this writing) have verified that our sample, taken as a whole, experienced large and moderate changes in psychiatric symptoms and interpersonal problems, respectively. We also assessed individual patient data using clinical significant change (Jacobson & Truax, 1991) and found that, following treatment, a sizable majority either recovered (38 %) or improved (31 %) when assessing symptomatic change. When assessing interpersonal problems, we found that 23 % recovered and 12 % improved. Most patients maintained these changes at the two-and-a-half-year follow-up. The second registration (neither published or completed as of this writing) is an analysis comparing patients with and without personality disorders (PD) in terms symptomatic and interpersonal improvements. The main finding from these preliminary analyses are that patients with and without a PD have the same level of improvements, but patients with PD are worse afflicted with psychopathology in general.

Hypotheses

There has been a considerable empirical investigation comparing the dose-response and GEL models (Robinson et al., 2019). Our aim is to supplement this literature utilizing a data set with significantly longer treatments and more severe psychopathology than has been investigated previously. We believe that our data will match previous estimates on rate of improvements were these overlap (sessions < 20), but we can only guess what the rate of improvements is for treatments that are longer than this. *Our main hypothesis follows the predictions from the GEL model, namely that there will be a linear relatiobship between treatment length and improvements made.*

Explanation of existing data

The original project was started in 1995, and data gathering was concluded in 2008. A number of scientific process-oriented publications have been published with results from the data set. See the link for a list of publications.
http://www.sv.uio.no/psi/forskning/prosjekter/multisenter/publikasjoner/index.html

Has the data been accessed?

See previously listed registrations. We have made deliberate attempts to secure that previous investigations do not influence our current hypotheses nor our choice of statistical procedures. The goal of this registration is to provide transparency into the statistical analyses of our data. Therefore, we feel the term “pre-registration” is misleading and apply a more conservative “registration”, to indicate that we have accessed the data before writing completing this registration.

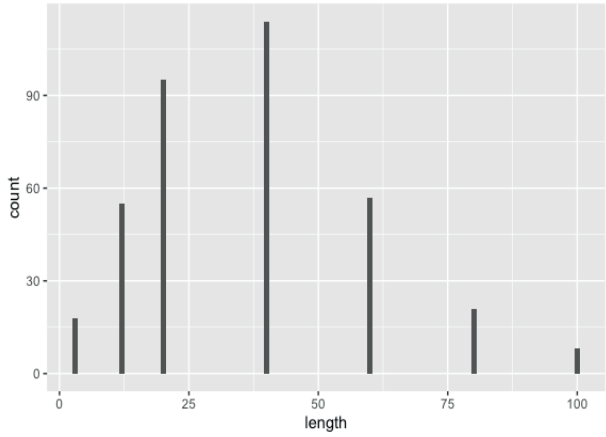
Data collection procedures. The patients were selected based on eligibility from a pool of regularly referred outpatients, and participation was based on informed and signed consent. We only used data from routine outpatient facilities in our analysis. We collected the majority of patient (n = 301) data from psychiatric outpatient clinics spread across 17 separate clinics. The clinics are nested in six over-arching treatment sites, which correspond to different hospital regions in Norway. We also gathered data from the Norwegian University of Science and Technology`s student clinic (n = 27). Lastly, we gathered data from an outpatient clinic (n = 42) specializing in psychodynamic body therapy for patient with somatoform disorders. Participant enrollment lasted from May 1995, throughout December 2000. Treatment was completed for every patient between the period of March 1996 to March 2007. At each of the eight sites, a specially trained coordinator (clinical psychologist or psychiatrist) was responsible for recruiting patients. The coordinators were instructed to select randomly from their local patient population but to ensure that roughly half had a personality disorder. The same coordinators carried out the assessment of the patients. Patients with serious substance abuse, psychosis, and acute crises requiring hospitalization as well as those with an IQ below 70 and who were younger than 20 years were not included in the study. The patients represented a heterogeneous and typical outpatient sample. Patients were not compensated for participating in the study.

Sample size. A total of N = 370 is available for analyses in the NMSPOP data set. As far as we are aware, this is the first psychotherapy trial that can examine rates of change in treatments that far exceeds the typical dosage amount (< 15) seen in Falkenstom et al., (2016) and Baldwin et al. (2009). This means that no comparison or simulation can be done to assess power by comparisons to other studies.
Measured variables. To assess symptomatic change, we will utilize the Symptom Checklist 90: General Severity Index (GSI). The GSI is the mean score across every item in the SCL-90. To assess changes in interpersonal functioning we will utilize the Inventory of Interpersonal Problems 64: IIP Global. The IIP Global is the mean score across every item in the IIP 64. Time will be measured using session number.
Indices
Clinical Significance. We will use Jacobsen and Truax (1984) definition of clinical significance to assess therapeutic gains. Clinically significant change occurs when a patient moves from a dysfunctional population to a functional or normal population during treatment and the magnitude of that patient's change is statistically reliable. A patient whose improvement meets both of these criteria is classified as recovered (having returned to normal functioning). On the basis of these criteria, patients are categorized as (1) recovered, (2) reliably improved but not recovered, (3) unchanged, or (4) deteriorated, in the case of reliable negative change. We plan on using predicted values from our multilevel regression models to assess clinical significance. The dependent variables will be SCL-90: GSI and IIP-64 Global. Each will be analyzed separately using linear mixed-effect models to estimate patient trajectories. We plan on replicating Falkenstrom et al. (2016) procedure. Each outcome variable will be assessed by two different statistical models which can be compared using model fit statistics. We plan on supplying both the Bayensian and Akaike Information Criteria.
Model 1: Dose Effect Model. In order to model the dose-effect hypothesis we will regress our outcome variables using a multilevel, third degree polynomial regression. This will enable an assessment of possible non-linear change during psychotherapy. The model will include a random intercept parameter to handle longitudinal dependency. We will apply Maximum Likelihood to focus on accurately estimated fixed effects.
Model 2: GEL Model. In order to model the GEL hypothesis, we will regress our outcome variables using a multilevel, third degree polynomial with an treatment length interaction term. This model will capture whether the change during treatment is related to the length of treatment as predicted by the GEL model. The model will also include a random intercept parameter to handle longitudinal dependency. We will apply Maximum Likelihood to focus on accurately estimated fixed effects.
Missing self-report data
Linear mixed-effect models are equipped to handle missing data as the models will include every relevant data when assessing individuals as a separate measurement level (Tasca & Gallop, 2009). We do not plan to use any other tools to handle missing data.
Analysis using simulated data

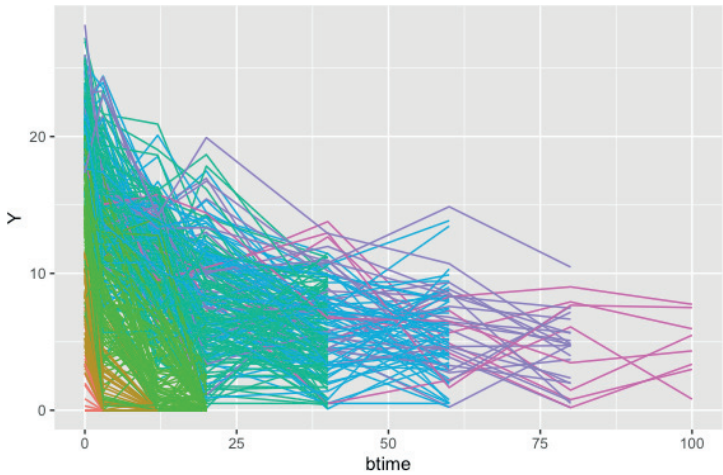
```
library(simstudy)
library(tidyverse)
library(data.table)

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose

library(lme4)

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

library(lmerTest)

##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step

library(ggeffects)

set.seed(555)
maxTime <- 80 # limit follow-up time to 80
def1 <- defData(varname = "nCount", dist = "noZeroPoisson",
                formula = 20)
def1 <- defData(def1, varname = "mInterval", dist = "nonrandom",
```

```r
                formula = 20)
def1 <- defData(def1, varname = "vInterval", dist = "nonrandom",
                formula = 0.4)
set.seed(20190101)
dt <- genData(370, def1)
dtPeriod <- addPeriods(dt)
dtPeriod <- dtPeriod[time <= maxTime]
def2 <- defDataAdd(varname = "mu", dist = "nonrandom",
                   formula = "20 - (1/500) * (time) * (180 - time)")
def2 <- defDataAdd(def2, varname = "var", dist = "nonrandom", formula = 9)
dtY <- genData(1000)
dtY <- addPeriods(dtY, nPeriod = (maxTime + 1) )
setnames(dtY, "period", "time")
dtY <- addColumns(def2, dtY)
dtY <- addCorGen(dtOld = dtY, idvar = "id", nvars = (maxTime + 1),
                 rho = .4, corstr = "ar1", dist = "normal",
                 param1 = "mu", param2 = "var", cnames = "Y")
dtY[, `:=`(timeID = NULL, var = NULL, mu = NULL)]
setkey(dtY, id, time)
setkey(dtPeriod, id, time)
finalDT <- mergeData(dtY, dtPeriod, idvars = c("id", "time"))
rm(maxTime,dtY,dtPeriod,dt,def1,def2)
sim <- finalDT
rm(finalDT)
sim <- sim %>%
  mutate(btime = case_when(period == 0 ~ 0,
                           period == 1 ~ 3,
                           period == 2 ~ 12,
                           period == 3 ~ 20,
                           period == 4 ~ 40,
                           period == 5 ~ 60,
                           period == 6 ~ 80,
                           period == 7 ~ 100,
                           period == 8 ~ 120,
                           period == 9 ~ 140,
                           period == 10 ~ 160,
                           period == 11 ~ 180,
                           period == 12 ~ 200,
                           period == 13 ~ 220,
                           period == 14 ~ 240,
                           period == 15 ~ 260,
                           period == 16 ~ 280,
                           period == 17 ~ 300,
                           TRUE ~ 999))
sim <- sim %>% group_by(id) %>% mutate(length = last(btime))
sim <- sim %>% select(id,btime,length,Y)
sim <- sim %>% ungroup()
# Replace below zero values with 0.1 or 0.3 or 0.5
sim <- sim %>% mutate(Y = ifelse(Y < 0,sample(c(0.1,0.3,0.5),1),Y))
# Remove patients that have less than n = 5
sim <- sim %>% filter(length %in% 3:100)

#

sim <- sim %>% mutate(Y = ifelse(length == 3,Y - 18,
                                 ifelse(length == 12, Y - 13,
                                        ifelse(length == 20,Y - 5,Y))))

sim <- sim %>% mutate(Y = ifelse(Y < 0,0,Y))

sim %>% filter(btime == 0) %>% group_by(length) %>% summarise(mean = mean(Y))

## # A tibble: 7 x 2
##   length  mean
##    <dbl> <dbl>
## 1      3  2.33
## 2     12  7.51
## 3     20 14.7
## 4     40 19.7
## 5     60 20.0
## 6     80 20.5
## 7    100 20.3
```

```
#


# Look at the data
sim %>% filter(btime == 0) %>% ggplot(.,aes(length)) + geom_histogram(binwidth = 1)
```



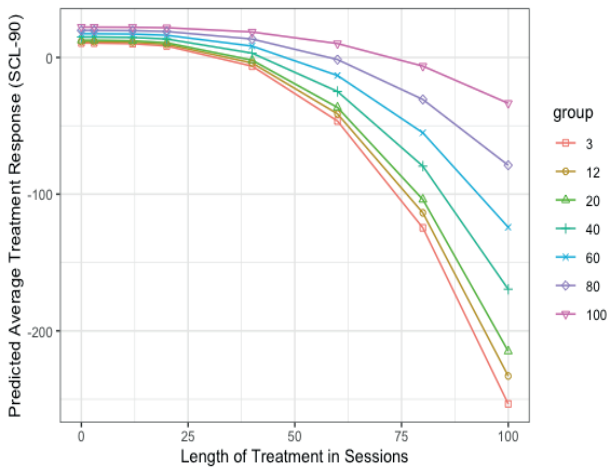

```
sim %>% filter(btime == 0) %>% summarise(mean = mean(length))

## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  35.5

sim %>% filter(btime == 0) %>% select(length) %>% table() %>% as_tibble()

## # A tibble: 7 x 2
##   .         n
##   <chr> <int>
## 1 3        18
## 2 12       55
## 3 20       95
## 4 40      114
## 5 60       57
## 6 80       21
## 7 100       8

ggplot(sim,aes(btime,Y,group = id)) +
  geom_line(aes(color = as.factor(length)))+
  #geom_smooth(method = "lm", inherit.aes = F,aes(time,Y)) +
  theme(legend.position = "none")
```




```
# Model data with GEL and Log-Linear models

# Make cubic sessions
sim$btime2 <- sim$btime^2
```

```r
sim$btime3 <- sim$btime^3

# Cubic growth model with random intercepts parameters
mod1 <- lmer(Y ~ btime + btime2 + btime3 +
               length +
               length*btime + length*btime2 + length*btime3 + (1 | id),
             data = sim,REML = F, na.action = na.exclude)

## Warning: Some predictor variables are on very different scales: consider
## rescaling

## Warning: Some predictor variables are on very different scales: consider
## rescaling

summary(mod1)

## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula: Y ~ btime + btime2 + btime3 + length + length * btime + length *
##     btime2 + length * btime3 + (1 | id)
##    Data: sim
##
##      AIC      BIC   logLik deviance df.resid
##   9462.9   9517.3  -4721.5   9442.9     1694
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.78181 -0.63472  0.00453  0.60954  2.86898
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept)  7.242   2.691
##  Residual             11.119   3.335
## Number of obs: 1704, groups:  id, 368
##
## Fixed effects:
##                 Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)    7.864e+00  3.718e-01  6.775e+02  21.148  < 2e-16 ***
## btime         -1.062e+00  4.800e-02  1.347e+03 -22.126  < 2e-16 ***
## btime2         3.104e-02  2.120e-03  1.347e+03  14.641  < 2e-16 ***
## btime3        -2.704e-04  2.583e-05  1.357e+03 -10.469  < 2e-16 ***
## length         1.919e-01  8.741e-03  6.456e+02  21.955  < 2e-16 ***
## btime:length   3.246e-03  8.827e-04  1.344e+03   3.678 0.000244 ***
## btime2:length -1.979e-04  2.825e-05  1.345e+03  -7.006 3.87e-12 ***
## btime3:length  2.147e-06  2.716e-07  1.357e+03   7.904 5.53e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) btime  btime2 btime3 length btm:ln btm2:l
## btime       -0.454
## btime2       0.279 -0.868
## btime3      -0.168  0.666 -0.924
## length      -0.842  0.316 -0.136  0.036
## btime:lngth  0.439 -0.802  0.516 -0.260 -0.469
## btim2:lngth -0.344  0.877 -0.816  0.604  0.310 -0.867
## btim3:lngth  0.246 -0.798  0.944 -0.903 -0.166  0.581 -0.871
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling

# Plot marginal effects
p <- ggpredict(mod1,terms = c("btime3","length"),type = "fe")

p$btime <- dplyr::recode(p$x, `0` = 0L, `1` = 1L,`27` = 3L,`64` = 4L,
                           `1728` = 12L,`2197` = 13L, `8000` = 20L,`9261` = 21L,
                           `19683` = 27L,`64000` = 40L,`68921` = 41L,`216000` = 60L,`226981` = 61L,
                           `512000` = 80L, `531441` = 81L,`1000000` = 100L, `1030301` = 101L)


ggplot(p,aes(btime,predicted,color = group)) +
  geom_line() +
  geom_point(aes(shape = group)) +
  scale_shape_manual(values = 0:7) +
```

```
  xlab("Length of Treatment in Sessions") +
  ylab("Predicted Average Treatment Response (SCL-90)") +
  theme_bw()
```




```
mod2 <- lmer(Y ~ btime + btime2 + btime3 +
                (1 | id),
             data = sim,REML = F, na.action = na.exclude)

## Warning: Some predictor variables are on very different scales: consider
## rescaling

## Warning: Some predictor variables are on very different scales: consider
## rescaling

anova(mod1,mod2)

## Data: sim
## Models:
## mod2: Y ~ btime + btime2 + btime3 + (1 | id)
## mod1: Y ~ btime + btime2 + btime3 + length + length * btime + length *
## mod1:     btime2 + length * btime3 + (1 | id)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod2  6 9906.4 9939.0 -4947.2   9894.4
## mod1 10 9462.9 9517.3 -4721.5   9442.9 451.45      4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

- Anderson, E. M., & Lambert, M. J. (2001). A survival analysis of clinically significant change in outpatient psychotherapy. *Journal of Clinical Psychology*, 57(7), 875–888. doi:10.1002/jclp.1056
- Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose–effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, 77(2), 203-211. Draper, M. R., Jennings, J., Baron, A., Erdur, O., & Shankar, L. (2002). Time-limited counseling outcome in a nationwide college counseling center sample. *Journal of College Counseling*, 5(1), 26–38. doi:10.1002/j.2161-1882.2002.tb00204.x
- Falkenström, F., Josefsson, A., Berggren, T., & Holmqvist, R. (2016). How much therapy is enough? Comparing dose-effect and good-enough models in two different settings. *Psychotherapy*, 53(1), 130-139. http://dx.doi.org.ezproxy.uio.no/10.1037/pst0000039
- Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist*, 41(2), 159-164. doi:10.1037/0003-066X.41.2.159
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19. https://doi.org/10.1037/0022-006X.59.1.12
- Robinson, L., Delgadillo, J., & Kellett S., (2019) The dose-response effect in routinely delivered psychological therapies: A systematic review, *Psychotherapy Research*, DOI: 10.1080/10503307.2019.1566676
- Tasca, G. A., & Gallop, R. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. *Psychotherapy Research*, 19(4-5), 429-437. https://doi.org/10.1080/10503300802641444