A Longitudinal Analysis of the Performance of Evidence-Based Therapists

Sarah Coe-Odess

B.A., Swarthmore College 2015

M.A., University of Virginia, 2017

A Dissertation Defense Presented to the

Graduate Faculty of the University of Virginia

for the Degree of Doctor of Philosophy

Department of Psychology

University of Virginia

June 16, 2022

Committee Members:

Bethany A. Teachman, Ph.D. (Chair)

Cannon Thomas, Ph.D. (Chair)

Patricia Lee Llewellyn, Ph.D.

Xin Tong, Ph.D.

J. Kim Penberthy, Ph.D., ABPP

Table of Contents

Abstract 4

A Longitudinal Analysis of the Performance of Evidence-Based Therapists……..…...7

Early Termination.……..…..…….....10

Improvement in Therapist Process.……..…….....11

Predictors of Patient Outcomes.……....13

Dissertation Overview……..…….16

Study 1……...……..……..16

Study 1a Method……..……..18

Treatment...……...……...…..18

Procedure……...……18

Participants……18

Measures……20

Study 1a Data Analysis Plan…….……..……...23

Study 1a Results…….27

Study 1a Discussion……..…….31

Study 1b Method……..……..32

Participants……...…..32

Measures……33

Study 1b Data Analysis Plan…….……..……...34

Study 1b Results……...…….34

Study 1b Discussion……..…….37

Study 1 Discussion……..…...……...…….38

Study 2 Method……..41

Participants……...…..42

Study 2 Data Analysis Plan…….……...43

Study 2 Results……..44

Study 2 Discussion…….48

General Discussion……....49

Limitations…….50

References……..……53

Abstract

Do professionals become more effective at their jobs with experience? Although we tend to assume more experienced professionals are more effective, research has not consistently found this to be the case in medicine and in psychotherapy (e.g., Burns & Wholey, 1991; Flood, Scott, Ewy, & Forrest Jr, 1982; Okiishi, Lambert, Nielsen, & Ogles, 2003; Wampold & Brown, 2005). Most research to date examining this question, however, has used a cross-sectional design (e.g., Budge et al., 2013; Okiishi et al., 2003; Okiishi et al., 2006; Wampold & Brown, 2005), focusing on differences in the performance of groups of therapists with different levels of experience. To date, there have been only two longitudinal studies that investigated directly whether therapists improved with time and experience. Both of them collected data from therapists who had an average of less than five years of data in the dataset, and therapists included both individuals in training and licensed professionals (Goldberg, Babins-Wagner, et al., 2016; Goldberg, Rousmaniere, et al., 2016). Despite this, the study that had a clearer system of informative feedback suggested that therapists can improve with experience, whereas the other study did not. This dissertation will focus on evaluating whether evidence-based therapists in a private practice setting do, in fact, improve their performance on various patient outcomes with experience/time, as well as whether other elements of their practice (specifically, decision-support tools that represent 'best practices' in providing evidence-based care) improve over time. These questions will be evaluated across two studies.

Study 1 is a single-therapist case study, in which we analyze trajectories of change for one therapist who has collected session-by-session data in her private practice across 39 years. Study 1 is idiographic in nature. We are particularly interested in whether a therapist who has collected data since her first year as a licensed clinician and has dedicated her career to evidence-

based practice has experienced improvement in outcomes over time. Measurements of improvement in patient outcomes include changes in degree of improvement on standardized outcome measures, efficiency (speed) of achieving desired outcomes, and changes in the reported quality of her terminations. Measurements of improvement in elements of her practice include changes in the consistency of using decision-support tools—including plotting patient therapy outcomes, writing treatment goals, creating case conceptualizations—and facilitating collaborative terminations. We are also looking at whether the therapist's caseload and the number and intensity of her personal and professional life events during a given year predicted her patients' outcomes during that year. Finally, we will examine whether her ability to use decision-support tools effectively and to prevent life events from negatively impacting her patient outcomes improves with experience. Results of Study 1 will be used to inform Study 2.

Study 2 will consist of data from up to 16 therapists in private practice and 674 of their patients who, at some point between 1995-2007, worked at the San Francisco Bay Area Center for Cognitive Therapy. Study 2 will build off of Study 1 by examining whether patterns of change tied to patient outcomes seen in Study 1 generalize to other therapists at the center or whether specific groups of therapists show different patterns of change.

Given the small sample, we cannot make conclusions on whether results of Studies 1 and 2 will generalize to broader groups of therapists and psychotherapy settings. Results of this dissertation, however, highlight the value of an idiographic approach to: (a) examine whether a therapist becomes more effective at reducing patient symptoms and achieving collaborative terminations with time and experience (providing a model that can be used for other therapists and in settings where clinics want to understand trajectories of change for their providers), (b) examine improvement over time in therapeutic practices; specifically, use of decision-support

tools, (c) evaluate other factors that may influence patient outcomes, including therapist life events and characteristics of the therapist caseload, and e) consider how therapists' effective use of certain factors or ability to cope with events, such as decision-support tools and life events, respectively, might improve with time. Together, these results will enhance understanding of what factors may affect therapist effectiveness in real-world settings across the course of therapists' careers.

The notion that "practice makes perfect" is a proverb that individuals hear as they practice sports and instruments, engage in hobbies, and learn new subjects and languages. Indeed, for chess players (Chase & Simon, 1973; Shanteau, 1992), musical performers (Lehmann & Gruber, 2006), and athletes (Hodges et al., 2006), there is evidence to support the idea that individuals improve their performance with years of experience. In medicine and psychotherapy, however, research has not consistently found this to be the case (e.g., Burns & Wholey, 1991; Flood et al., 1982; Okiishi et al., 2003; Wampold & Brown, 2005).

Ericsson (e.g., 2006, 2007) has proposed that, in fields in which expertise does not develop, individuals often do not have a system in which they receive clear, informative feedback about their performance. This idea of a clear system of feedback, considered "deliberate practice" involves not only reviewing results but also having a laid-out protocol for how to adjust if patients are not progressing as expected. The importance of deliberate practice in therapy has been supported by others, who have found that, for therapists, simply executing the same skillset continually is insufficient for improvement (Chow et al., 2015), that experience as a therapist is not predictive of accuracy and skill (Tracey et al., 2014), and that experience and/or professional degree do not account for much variability in outcomes among therapists (Wampold & Brown, 2005). In fact, some naturalistic studies have found that therapist trainees yield similar outcomes to their licensed counterparts (Budge et al., 2013; Okiishi et al., 2003), and others have occasionally found small but significant effect sizes ($d=0.15$) that indicate that trainees yield better outcomes than professionals (Stein & Lambert, 1995). In one study of 92 therapists with an average of 16.5 ($SD=14.8$) patients, results suggested that therapists are able to utilize prior experience to help future outcomes only if the patients are similar and are treated within a short time span of one another (Leon et al., 2005).

Some researchers, however, believe that therapists *can* improve with experience. In a multinational study of more than 4,000 therapists, Orlinsky and Rønnestad (2005) found that therapists who have been practicing for at least 15 years are more likely than individuals who have been practicing for fewer than 15 years to be effective with patients. They proposed that the reason for this is that, over time, therapists are able to accumulate skill, increase their mastery, and surpass prior limitations in their clinical work. Skovholt et al. (1997) agreed that 10-15 years is a benchmark for expertise, as that amount of time allows for practitioners to comprehend theory and research, develop a working style, and create an accurate way to evaluate success.

Most existing research on changes in therapist effectiveness over time has been conducted using cross-sectional designs (e.g., Budge et al., 2013; Okiishi et al., 2003; Okiishi et al., 2006; Wampold & Brown, 2005). While informative, cross-sectional studies are limited in that they do not assess the course of change for individuals over time. To our knowledge, there have been only two longitudinal studies that investigated whether patients experienced greater symptom reduction as their therapists gained experience. These studies suggest that clear informative feedback is crucial to learning from experience. One study used data from 170 therapists with an average of 4.73 years of data per therapist and had a feedback system via supervision in which clinicians viewed selected video-recorded therapy sessions with supervisors each week; it did not, however, have a clear protocol for feedback and, thus, lacked an accountability mechanism for the supervisor to adhere to giving feedback and lacked a deliberate process for reconsidering the treatment plan and challenging established processes of therapy. This study found a small but significant effect indicating that therapists' outcomes slightly worsened over time (Goldberg, Rousmaniere, et al., 2016). The other study used data from 153

therapists with an average of 4.42 years of data per therapist, and clinicians engaged in a feedback-informed treatment model of case review (Miller et al., 2015), in which clinicians met monthly with an outside consultant about cases that were not progressing to make plans for adjusting treatment to meet the needs of patients. This study found that, on average, therapists did improve slightly over time (Goldberg, Babins-Wagner, et al., 2016). Consistent with Ericsson's (2006, 2007) theory of deliberate practice, it appears that an evidence-based setting with clear protocols for using information about outcomes to guide treatment provide conditions for improvement in performance over time that other settings lack.

Although the results of the latter study are promising, the current evidence is limited; given the importance of improvement in outcomes through training, we cannot assume that the conditions in one clinic were sufficient to show whether growth over time is possible. Furthermore, in Goldberg, Babins-Wagner et al.'s (2016) study, over half of the sample size of therapists were still students, further limiting the ability to make assumptions about improvement among licensed professionals. In the present study, we conduct an in-depth analysis of one evidence-based therapist who applied practices that established accountability and informative feedback throughout the course of her career and in her supervision of post-grad trainees. We use this idiographic approach to take a close look at how one psychologist has conducted her four-decade career; while results from a single therapist are clearly not generalizable to therapists in general, we believe the methods are generalizable and this approach provides an example for how other therapists could both draw from her heavily feedback-informed model of care and examine their trajectories over time.

To measure whether the primary therapist (Studies 1 a and b) and her trainees (Study 2) improved over time, we will use three dimensions of improvement: degree of patient symptom

reduction, speed of patient symptom reduction, and quality of terminations. To assess symptom reduction, we will be utilizing session-by-session data from a depression measure. Prior research (Goldberg, Babins-Wagner, et al., 2016; Goldberg, Rousmaniere, et al., 2016) has utilized pre-post data to assess effect sizes. While useful, session-by-session analyses provides a more fine-tuned analysis of the data, as they take into account the potential that one week's scores may not be an overall reflection of the entire course of treatment.

Early Termination.

Prior research has viewed dropout rate as an indicator of therapist effectiveness. One of the aforementioned longitudinal studies on whether therapists improve over time looked at whether rates of one- or two-session treatment durations decreased over time (Goldberg, Rousmaniere, et al., 2016). The authors found that, although therapists' degree of patient symptom improvement slightly worsened over time, therapists' rates of "early" terminations decreased over time. One reason that unplanned terminations might decrease with experience is that patient-therapist alignment on treatment plans and goals seems to predict fewer unplanned terminations (Tracey, 1986). Thus, therapists who work to engage in these collaborative processes throughout their careers might yield increasingly positive results.

Goldberg, Rousmaniere, et al.'s (2016) study did not have a way to measure whether these short treatment durations were planned or not. In other words, some treatments might be short in duration because treatment goals have been met, and other treatments might be longer in duration but end abruptly. In the present study, we consider uncollaborative terminations (whether the patient and therapist worked well together on the termination, discussed it fully, and agreed on it) and premature terminations (whether therapy was given a fair shake/tried for long enough to help patient accomplish their treatment goals). Uncollaborative dropout is particularly

important to measure because when patients do not end therapy collaboratively, therapists are not necessarily aware of how the patient felt about the course of therapy (Westmacott et al., 2010).

Some studies of dropout have found a significant therapist effect of 6.21-12.6% of the variance in the percentage of patients' dropping out prematurely after controlling for initial impairment (Swift & Greenberg, 2012; Zimmermann et al., 2017), indicating the modest importance of therapists in predicting and preventing dropout or early termination. Studies to date, however, have not operationalized early termination or dropout in a consistent fashion (Swift & Greenberg, 2012; Westmacott et al., 2010). Overall, various definitions—among other factors leading to discrepant outcomes across studies—have led to reported early termination rates between 20-60% (see Reneses et al., 2009). Both the degree of collaboration and whether the termination was premature have been used in these definitions, so we include measures of both here.

Improvement in Therapist Process.

Not only might therapists improve with respect to patients' symptom reduction outcomes or rates of collaborative terminations, but therapists' technique or process might also improve. The therapists in the present studies measured three treatment processes (plotting outcome measures, writing individualized case formulations, and recording treatment goals) throughout the duration of the dataset. The present studies will look at whether the therapists became more consistent in their use of these processes and whether the processes became more predictive of better outcomes over time, consistent with the idea that the tools are being used more effectively (though this is not directly measured).

It has been well-documented that systematic, immediate feedback regarding outcomes is crucial to improvement, as is the ability to implement changes in future situations by learning

from prior errors (Bickman, 1999; Ericsson, 2006; Tracey et al., 2014). In psychotherapy, this is often not possible, as therapists often do not have clear feedback on each course of treatment until termination occurs. Similar to what medical fields have adopted, routine outcome monitoring (ROM) occurs when objective measures of progress are completed by patients regularly, and therapists (and sometimes patients) have access to these outcomes (Howard et al., 1996; Lambert et al., 2001; Miller et al. , 2005; Shimokawa et al., 2010; Wampold, 2015). These data can be used both to compare a specific patient's outcomes to typical change trajectories (Lambert et al., 2005; Miller et al., 2005) and to detect, in real time, patients who are worsening in symptoms (Hannan et al., 2005).

Outcome monitoring is particularly effective when outcomes are utilized to provide feedback to the therapist and the patient in a way that holds the therapist accountable to outcomes and facilitates conversations with the patient about the progress or lack thereof. One way to do this is to plot a patient's outcomes over time, to show the trajectory of symptoms since therapy began. Research has demonstrated, both in randomized controlled trials (RCTs) and in naturalistic settings, that outcome monitoring can positively affect outcomes and reduce the rates of patients who are not progressing as expected (Lambert et al., 2002, 2003; Lambert & Shimokawa, 2011; Shimokawa et al., 2010; Whipple et al., 2003). It can also enable a faster treatment response (de Jong et al.,, 2012; Lambert & Shimokawa, 2011) and allow therapists to adjust treatment as needed to improve outcomes (Carlier et al., 2012; Harmon et al., 2007). Importantly, these results only hold true when patient outcomes are used to consider why patients are or are not progressing (Lambert et al., 2002, 2003).

Another technique on which therapists might improve is case formulations, in which therapists individualize treatment plans that guide what interventions to use for each particular

patient. According to Persons and Tompkins (1997), case formulations should include a list of patients' problems and diagnoses and a hypothesis about the cause of the problems' start and continuation, as well as the relationship between the problems and the diagnosed disorder(s). Compared to following a standardized protocol, conducting case formulation-driven psychotherapy allows for multiple diagnoses to be targeted simultaneously and for treatment to be responsive to the individual patient's needs and their progress (Persons & Hong, 2015).

In addition to improving their use of outcome monitoring and case formulations over time, some therapists may also become more consistent at recording individualized treatment goals. Treatment goals typically include both reduction of symptoms and any idiographic goals (e.g., strengthening relationships or completing projects or assignments; Persons et al., 2006). Collaboratively setting goals can help with buy-in and measuring treatment progress (Padesky & Greenberger, 2012; Shirk & Karver, 2003; Shirk & Saiz, 1992), as well as a sense of teamwork between the therapist and patient (Tryon & Winograd, 2011).

Given the importance of these tools in creating tailored courses of therapy that suit and respond to each patient's needs, one would expect that using them would improve patients' outcomes and facilitate a stronger therapeutic alliance. All of these decision-support tools are standard practice at the center from which data for the following studies were drawn. The current studies will assess: (1) whether a therapist's adherence to these best practices increases over time, and (2) whether the use of the tools becomes more closely associated with improved outcomes over time, which may indicate the clinician has become more skilled in utilizing them to improve treatment (though causal inferences about the impact of using the tools are not possible given the correlational nature of these data). Thus, these studies will use decision-

support tools as outcome measures in the main analyses, but we will also conduct secondary analyses using decision-support tools as a predictor of patient outcomes.

Predictors of Patient Outcomes.

In addition to changes in a therapist's effectiveness over time, there are numerous other factors that vary over time that could affect therapist performance. We focus on two important factors (size of caseload and therapists' professional and personal life events). If there is a relationship between these factors and patient outcomes, the results would have implications for therapists regarding management of these issues in order to optimize their psychotherapy performance. Research on the effects of caseload on therapist effectiveness has yielded mixed results. Some research has indicated that caseload negatively predicts therapist effectiveness (Vocisano et al., 2004), whereas other research has found that clinicians with higher caseload per clinic day tend to yield better patient depression and anxiety outcomes (Firth et al., 2015). One potential reason for this latter association might be that therapists with higher caseloads gain more treatment experience, thus aiding their clinical abilities (Firth et al., 2015), whereas the former association may indicate therapist burnout with caseloads that are too high. The current studies will explore whether higher caseload will predict concurrent outcomes, either positively or negatively.

Another factor that might impact a therapist's performance at any specific time is non-clinical life events or obligations. In examining the extended career of therapists, it is to be expected that various life events and obligations—both positive and negative—will occur. In a 1989 survey of 749 practicing therapists, 74.3% of therapists reported personal distress in the past three years, and, among those, 36.7% indicated the distress had impacted the quality of their work negatively (Guy et al., 1989).

Interestingly, different facets of therapists' life events and associated reactions may have variable effects. Nissen-Lie et al. (2013) designed a naturalistic study that included data from 16 outpatient therapy clinics in Norway. Therapists had a range of experience, with a mean of 10 years of professional experience and an average caseload of five patients per week. Using a measure that asked questions about therapists' level of life satisfaction and stress, among other factors, the authors found that therapists' stress levels negatively affected patients' reports of working alliance, but therapists' life satisfaction did not affect patients' reports of working alliance. Another survey of 552 therapists found that self-reported professional and personal distress were associated with self-reported work impairment. Therapists with more experience reported less work impairment originating from professional distress, but experience did not seem to mitigate the impact of personal distress on work impairment (Sherman & Thelen, 1998). The authors proposed that a potential reason for the former finding was that experienced therapists had developed better coping skills—or resilience—to deal with distress, although this theory does not explain the latter finding.

Other researchers have also identified resilience as a key trait in effective therapists (Green et al., 2014; Luborsky et al., 1985). One cross-sectional study of 21 Psychological Wellbeing Practitioners at six service sites measured practitioners' self-report of resilience—among other measured qualities—and found that individuals with higher self-reported levels of resilience were more likely to have better outcomes with their patients (Green et al., 2014). To our knowledge, no study to date has looked at whether therapists' life events predict their patients' outcomes. The current studies will examine whether a therapist's personal and professional life events and obligations—both positive and negative—predict patient outcomes in therapy and, if

so, whether this predictive power decreases over time, suggesting, perhaps, that therapists acquire resilience with experience.

Dissertation Overview.

The following two studies will seek to address the questions of whether and in what ways feedback-informed therapists improve over time, and whether factors outside of therapists' years of experience predict therapy outcomes, including symptom reduction and termination. Studies 1 a and b will focus on an evidence-based therapist with 40 years of data, who has focused her career on applying practices that established accountability and informative feedback throughout the course of her career. Study 2 will focus on psychologists whom she supervised at her Center during their first two post-graduate years.

Study 1

In the following two-part longitudinal, single-therapist case study, we will analyze trajectories of change for one therapist who has collected session-by-session data in her private practice across 40 years. The first study covers the first 26 years of her career and uses the Beck Depression Inventory (BDI), which was collected at every session, as the primary outcome measure. The second covers the last 14 years of the 40-year measurement period and uses the Depression Anxiety and Stress Scale (DASS), also collected at every session, as the primary outcome measure.

We are particularly interested in whether a therapist who has collected data since her first year as a licensed clinician and has dedicated her career to evidence-based practice has experienced improvement in outcomes over time, using a range of measures of improvement. For the dataset containing the first 26 years of her career, it was hypothesized that: (a) the therapist's patients would show greater symptom reduction in psychotherapy as the therapist

gained experience over time; (b) the therapist would experience fewer unilateral, uncollaborative terminations as the therapist gained experience over time; (c) the therapist would engage in more outcome monitoring, case conceptualization, and setting treatment goals over time (i.e., use of decision support tools); (d) the therapist would become more efficient as indicated by faster reduction of patient symptoms over time; and (e) amount and magnitude of life events and professional responsibilities in the therapist's life would negatively predict patients' concurrent outcomes. We also ran an exploratory analysis to assess whether (f) greater caseload would predict better or worse concurrent patient outcomes. We also predicted that, as the therapist gained experience, she would become more adept at utilizing her therapeutic skills. To assess this, we hypothesized that (g) the therapist's use of decision-support tools would increasingly predict better patient outcomes and fewer uncollaborative terminations over time, such that the use of a given support tool became more strongly related to symptom reduction over time. Finally, we hypothesized that (h) the relationship between life events and professional responsibilities on patient outcomes and quality of terminations would lessen as the therapist gained experience.

Given research that 10-15 years is the typical amount of time necessary for therapists to improve significantly (Orlinsky & Rønnestad, 2005; Skovholt et al., 1997), we expected that Study 1b would indicate an asymptote at the level the therapist achieved by the end of Study 1a. Thus, for the last 14 years of her career, it was hypothesized that: (a) the therapist's outcomes, as measured by patients' symptom reduction, would not continue to improve; (b) the therapist would maintain a low rate of uncollaborative terminations that she had achieved by the end of Study 1a; (c) the therapist's use of case formulation, plotting outcomes, and recording treatment goals would continue at the level she had achieved by the end of Study 1a; (d) greater caseload

would predict better concurrent patient outcomes, as measured by standardized end scores; and (e) amount and magnitude of professional life events in the therapist's life will negatively predict patients' concurrent outcomes, but personal life events will not.

Study 1a

Method

Treatment

Treatment consisted of individual cognitive behavioral therapy (CBT), typically provided weekly. The therapist did not usually adhere to a manualized treatment. Instead, she developed an individualized cognitive-behavioral case formulation for each patient and used the formulation and the results of progress monitoring to make treatment decisions, including to select interventions from available CBT manuals and other sources (Persons, 2012). Treatment was open-ended in duration and ended ideally when patient and therapist agreed that treatment goals had been reached.

Procedure

Patients completed an outcome measure in the waiting room before each session, including the intake sessions, and the therapist typically plotted the score at the beginning of the session, reviewed the plot with the patient, and used the data to guide the session. These patients were monitored with standardized outcome measures at each therapy session. The BDI was the primary measure used for session-by-session outcome monitoring. Variables regarding termination and decision-support tools were coded by the therapist at the time that the database was compiled, based on a review of the full clinical record and their recollections.

Participants

The data for Study 1a were drawn from an archival database comprised of information from 1472 adult patients who received individual psychotherapy during the years 1981-2008 from one of 20 licensed therapists or trainees operating in an evidence-based group practice between the years of 1995 and 2008 or in the founding director's individual practice prior to the inception of the group (1981-1994).

The present study will focus on the 536 patients treated by the founding director of the Center, either in her individual practice or at the Center. The first year of data coincided with her first year after her doctoral internship. The dataset includes all of her patients monitored with standardized measures in the first 26 years of her practice and represents most of the patients she treated during that time. The dataset also includes variables that detailed types of termination and use of decision-support tools (plotting outcomes, writing conceptualizations, recording treatment goals), among other variables. All patients gave written permission for the use of data that were collected as part of their treatment for research purposes.

Patients were included in the present study if they: (a) had BDI scores of 10 or higher during their first and/or second therapy session, (b) had not specified prior to attending a meeting that the purpose of the meeting was consultation only, and (c) had more than one therapy session. A cutoff score of 10 was used because this is the threshold for mild depression on the BDI. To simplify data analysis, when patients had more than one course of treatment within the dataset, we included only their first course.

The sample that met these criteria consisted of 218 of 536 (40.67%) patients. The average patient in the sample was 39.35 years of age (*SD* = 14.34) and had completed 16.66 years of education (*SD* = 2.76). Information about participants' racial and ethnic identities is limited in that race and ethnicity were combined into a single question in the original dataset.

Most of the sample (83.49%) identified as Caucasian, 5.05% as Asian, 2.75% as Hispanic, 2.75% as African American, and 1.38% as other. The remaining 4.58% did not have a recorded race or ethnicity on file. Most of the sample (60.09%) identified as female. Most patients (86.70%) were diagnosed with at least one anxiety or depressive disorder. Psychiatric diagnoses were made based on a clinical interview by the therapist, who used the most current version of the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1987, 1994, 2000) available at the time the patient was treated. The median first BDI score of the sample used was 19 (*SD* = 7.99). About half of patients (57.34%) received adjunctive pharmacotherapy at some point during treatment and some (24.77%) received adjunctive psychosocial treatment (e.g., couples therapy, Alcoholics Anonymous).

Measures

Patient Outcomes

Beck Depression Inventory. Symptoms of depression were measured with the original version of the BDI (Beck et al., 1961). The BDI is a widely-used, 21-item self-report measure of the severity of depressive symptoms that has been shown to have good internal consistency (α = 0.86 for psychiatric patients) and good convergence with other measures of depressive symptoms (Beck et al., 1988).

Uncollaborative termination. Uncollaborative termination was coded positively if the therapist answered "no" to the question: "Did the patient and therapist work well together on the termination, agree on it, and discuss it fully?" The termination was coded as uncollaborative, for example, if the patient cancelled a session and never rescheduled.

Premature termination. Premature termination was coded positively if the therapist answered "no" to the question: "Has the therapy been given a fair shake/tried for long enough to

help patient accomplish [their] treatment goals?" Reasons for the therapist's judgment that termination was premature varied from case to case, depending on the patient's treatment goals, which varied extensively and ranged from "reducing scores on the BDI to the normal range" to "beginning to date and finding a boyfriend." Uncollaborative and premature were not mutually exclusive: termination could be coded as both premature and uncollaborative.

Decision-Support Tools.

Plotting outcomes. Plotted outcome was coded positively if there was a plot of BDI scores in the chart with at least one score entered.

Case formulation. Case formulation was coded positively if there was a written case formulation in the clinical chart that was separate from the intake note or any other clinical note. In order to be counted, there needed to be a written formulation of the entire case, not just a brief mini-formulation that only focuses on one problem or symptom (e.g., a Thought Record or diagram about just one problem).

Treatment goals. Treatment goals was coded positively if the goals of treatment were stated in the clinical chart prior to the termination note. These were the goals of treatment developed and specified in collaboration with the therapist, rather than the general desires for change that brought the patient into the first session of treatment. To count, the word "Goals" or "Objectives" must have appeared, and there had to be a list, except in rare cases where there was a single goal. It was not sufficient to state, "The patient seeks treatment to work on OCD symptoms" or something similar.

Predictors of Patient Outcomes.

Caseload. Caseload was measured by the number of therapy sessions that occurred in a given year. Because of inconsistent frequency in attending therapy, measuring number of

sessions is a better measurement than number of patients in assessing clinical workload. All therapy sessions—not just the ones that met inclusion criteria—were included in measurement of caseload.

Number of diagnoses. Number of diagnoses was measured by the total number of Axis I diagnoses given to each patient. Diagnoses were based on the therapist's judgments and were not consistently based on standardized diagnostic interviews. Number of diagnoses ranged from zero to five, with a mean of 1.80 diagnoses per patient.

Presence of an Axis II diagnosis. Presence of an Axis II diagnosis was measured by whether the patient was diagnosed with an Axis II disorder at intake. Recording multiaxial diagnoses was standard practice throughout treatment.

Presence of a depressive disorder. Presence of depressive disorder was measured by whether the patient was given a diagnosis categorized as a depressive disorder in the version of the DSM in use at the time of treatment at intake.

Life events inventory. A semi-structured video interview with the therapist was conducted via Zoom during the Spring of 2020. The interview took place over the course of three separate days. The therapist was asked to list every significant personal life event, personal obligations, professional life event, and professional obligations that she experienced between 1981-2019. Examples of personal life events included moving across the country or giving birth. Examples of personal obligations included helping her daughter buy a house or changing her work schedule to take care of her grandson. Examples of professional life events included having a patient with multiple suicide attempts or starting a new group practice. Examples of professional obligations included being the editor of a journal or writing a book. The therapist then rated how exciting/engaging, stressful, and time-consuming each event/obligation was on a

scale from 0 Negligible (Hardly at all) to 3 High (Dominant). Anything coded as a 2 or 3 was considered "moderate or higher." For time-consuming events with moderate or higher scores, the therapist also indicated for how long the event remained at least moderately time-consuming. For events that remained time-consuming across multiple years, the event was counted for each year. Stressful and exciting/engaging events were not mutually exclusive: it was possible for an event to receive a score of 3 on stress level and on excitement/engagement level.

Data Analysis Plan

The average course of treatment lasted 33.84 weeks (SD = 56.97, median = 15). To assess changes in BDI scores during treatment, we ran mixed-model growth curve models using up to the first 20 weeks of therapy, regardless of the number of sessions. The heterogeneity of courses of treatment in private practice makes creating a standard model that generalizes across patients very difficult. Many patients remain in treatment for years, pursuing a range of treatment goals, and they can have no symptoms or go through multiple depressive episodes during that time. In the first 20 weeks of therapy, however, the majority of patients at this therapist's clinic shared the goal of reducing their initially high depressive symptoms. By restricting our analyses to the first 20 weeks, we were looking at a stage of treatment where there is clinically a relatively high degree of commonality. This decision was supported by preliminary analyses showing:(1) the majority of change in therapy occurred within the first 20 weeks of treatment for the vast majority of patients and (2) that linear mixed model growth curves provided adequate fit for the first 20 weeks but very poor model fit if the full course of treatment was used.

In naturalistic settings, it is typical to have substantial variability in the frequency of attending therapy sessions (Erekson et al., 2015; Gutner et al., 2016). In an effort to determine whether it is best to understand progress in treatment based on length of time in treatment,

number of sessions completed, or both, we calculated frequency of sessions by dividing the number of therapy sessions within the first 20 weeks by the number of weeks of treatment (with 20 weeks as the maximum). Frequency ranged from approximately three times per week to once every three weeks with an average frequency of 0.99. To account for the possibility that number of therapy sessions might be a better indicator for a cutoff than number of weeks since therapy began, we created a multi-level model, with number of weeks, frequency of sessions, and an interaction term between the two as predictors of end BDI score. Frequency of session was not significant, but number of weeks was; thus, subsequent analyses continued to focus on length in terms of weeks rather than sessions.

Of course, these choices come with tradeoffs; while making sense for the group of patients as a whole, the decision to use only the first 20 weeks likely was not optimal to evaluate degree and rate of change for many patients at the individual level, given the inherent variability in patients' treatment course. This 20-week timespan did not preclude use of the termination variables: terminations that occurred after 20 weeks of therapy were still included in termination analyses. Overall, this option allowed us to use mixed model growth curve modeling, which would not have been possible without fixing the number of weeks to be constant across patients.

To determine psychotherapy outcomes, we used the output from the above model to estimate the change in BDI score for each patient at the end of treatment. We did not simply use change from initial BDI to final BDI as a measure of improvement because: (a) patients who start with a higher initial severity in general improve more in psychotherapy; (b) a significant proportion of the patients in this sample are missing an initial BDI; and (c) taking advantage of all of the data available for each patient allowed us to produce a more reliable estimate that is less affected by a single weekly fluctuation in scores. Using the parameters estimated from a

mixed model growth curve allowed us to produce a regression equation for each patient that can be used to calculate a patient's "true" score at the last time point in therapy. These mixed model growth curves produced a fixed intercept and a fixed slope for the sample, and random intercepts and random slopes that are specific to a particular patient. The sum of the fixed (central tendency across the sample) and random (patient specific deviation from the fixed value) intercepts produced an intercept, and the sum of the fixed and random slopes produced a patient specific slope. The resulting regression line allowed prediction at any point during the patient's psychotherapy, and we used it to predict the expected score at the last point in their first 20 weeks of treatment.

Because average initial BDI scores may vary from year to year, we adjusted the equations so that predicted end BDI scores are the end scores that would be expected if all patients had the same level of symptom severity at intake. This allowed for an apples-to-apples comparison across groups. To achieve this result, we adjusted the slopes for each patient-specific regression equation to be the slope that would have been expected if the patient's intercept (initial BDI) had been equal to the fixed intercept (the typical value across the entire sample). Mixed model growth curves produced an estimate of the correlation between the slope and the intercept of the model. In psychotherapy data, this correlation is almost always negative: Patients with higher initial scores (higher intercepts) improve faster (have a steeper negative slope). We used this correlation to adjust the slope to what would be expected if the patient's intercept had been equal to the fixed intercept:

$$(1)\ adjusted\ slope = unadjusted\ slope + random\ intercept * \left(slope.intercept\ correlation * \frac{standard\ deviation\ of\ the\ slope}{standard\ deviation\ of\ the\ intercept}\right)$$

We then predicted where the score would be at the "end" of each patient's treatment (at the last session within the 20-week period) based on total weeks from the beginning to the end of the maximum 20 weeks of treatment, using a regression equation with the fixed intercept as the intercept and the adjusted slope as the slope. We refer to this as the adjusted predicted end score for each patient.

We tested our hypotheses using linear regression analyses. All analyses were completed in R (RCore, 2016). Mixed-model growth curve models were conducted using the lme4 package (Bates et al., 2019). To account for the fact that patients' diagnostic profiles might predict degree of treatment, three measures of symptom severity—presence of a depression disorder diagnosis at intake, presence of an Axis II disorder diagnosis at intake, and total number of psychiatric diagnoses at intake—were entered simultaneously into a linear regression model predicting patients' adjusted predicted end BDI scores. Only presence of an Axis II diagnosis at intake was significant; thus, only presence of an Axis II diagnosis was controlled for in the subsequent regression model.

Therapist experience was computed based on the number of years of post-licensure clinical experience. Her first year of post-licensure clinical work was defined as year 0 (baseline). An initial analysis assessed therapist experience as a predictor of BDI outcomes, accounting for presence of an Axis II diagnosis, mean frequency of sessions in the first 20 weeks, and mean length of treatment up to the first 20 weeks. This allowed us to account for potential dose effects and assess not only whether duration in between sessions but also duration between sessions relative to total number of sessions accounted for some change in patient outcomes. Presence of an Axis II disorder was no longer a significant predictor of adjusted end score and was, thus, removed from subsequent models. This finding was consistent with prior

research that has indicated that patient factors beyond initial symptom score predict symptom scores at the end of treatment (Hansen et al., 2002). No future models contained control variables unless otherwise specified in analyses. Models that predicted outcomes other than end BDI score (i.e., nature of termination, use of decision-support tools) will be assessed via logistic regressions. For all linear regressions involving interaction terms, we mean-centered predictor variables. For any non-significant linear effects, we also checked if there was a significant non-linear effect.

Results

Therapist Experience as a Predictor of Patient Outcomes

Hypothesis: The therapist's patients would show greater symptom reduction in psychotherapy as the therapist gained experience over time. We are most interested in the proportion of variance that can be attributed to the therapist, or, in other words, whether we're able to predict her performance during any given year. This is most directly measured by the average adjusted end score of her patients during each year, since running the model with individual patients' scores focuses more on patients' performance, rather than the therapist's performance; there are a lot of patient influences that may hinder our ability to predict outcome, and we wanted to avoid having those skew the perception of the therapist's average performance in that year. Because of this, we created a model with years of experience predicting patients' *mean* predicted, adjusted end BDI score for that year. Our model indicated that the therapist did improve over time (β = -0.03, p= 0.04; see Table 1). Because running the model with individual scores is more consistent with a standard approach and because of the small N-value each year, we ran another model that predicted individual adjusted end BDI scores. The rate of improvement per year was the same, but the finding was not significant (β = -0.03, p= 0.49).

Hypothesis: The therapist would experience fewer unilateral, uncollaborative terminations as the therapist gained experience over time. We conducted logistic regressions to assess whether rates of collaborative vs. uncollaborative and pre-mature vs. not premature terminations changed as the therapist gained experience. There was a positive correlation between rate of uncollaborative termination and rate of premature termination ($r = 0.42$, $p < 0.001$). At baseline (in 1981), her probability of uncollaborative terminations was 44.53%, and her probability of premature terminations was 45.38%. It was found that the odds of yielding an uncollaborative termination decreased by 5.31% (95% CI [0.01, 0.10]) for each additional year of experience (see Table 2 for β and p-value). It was found that the odds of yielding a premature termination increased by 0.38% (95% CI [0.004, 0.01]) for each additional year of experience, although this finding was not significant (see Table 2 for β and p-value).

Hypothesis: The therapist would become more efficient as indicated by faster reduction of patient symptoms over time. To test the hypothesis that the therapist achieved faster reduction of patient symptoms over time, we ran a regression in which years of experience predicted the adjusted, predicted slope of improvement. Our final model indicated that the therapist did not become more efficient as she gained experience ($\beta = -0.005$, $p = 0.44$; see Table 2).

Therapist Experience as a Predictor of Therapist Process

Hypothesis: The therapist would engage in more use of decision-support tools over time. We conducted logistic regressions to assess whether the therapist's use of decision-support tools increased as she gained experience. At baseline, her probability of plotting outcomes was 58.13%, of recording treatment goals was 0.18%, and of writing individualized case formulations was 86.56%. The odds of plotting outcomes for each patient increased by 9.30% (95% CI [0.04,

0.15]) for every additional year of experience. There was insufficient variability to calculate the odds of recording treatment goals for every additional year of experience. The odds of writing individualized case conceptualizations decreased by 1.78% each year (95% CI [-0.07, 0.04]), although this finding was not significant (see Table 2 for β and p-values). There was, however, a significant quadratic effect of written case formulations, indicating that the therapist utilized case formulations most frequently toward the middle of her career (β = -0.02, p< 0.001).

Use of Decision-Support Tools as a Predictor of Patient Outcomes Over Time

Hypothesis: The therapist's use of decision-support tools would increasingly predict fewer uncollaborative terminations and more symptom reduction over time. To test the hypothesis that the therapist's use of decision-support tools would more strongly predict better patient outcomes and fewer uncollaborative terminations over time (suggesting the therapist became better able to utilize decision-support tools over time to yield better results), we ran six linear regressions: each with one of the mean-centered decision-support tools, the mean-centered year variable, and an interaction between the mean-centered tool and year as predictors and with either adjusted, predicted end BDI score or presence of a collaborative termination as an outcome. These analyses used individual adjusted end scores, rather than the mean per year, so that a patient's end score and the therapist's use of a decision-support tool for that specific patient would be associated with each other. As predicted, the therapist's use of plotting outcomes increasingly predicted lower adjusted end scores over time (β = -0.48, p= 0.002) and more collaborative terminations over time (β = 0.03, p= 0.02; see Figure 1). No other models indicated significant effects.

Personal and Professional Factors as Predictors of Patient Outcomes

Hypothesis: Greater caseload would predict concurrent patient outcomes. To test the hypothesis that the therapist's patient outcomes would vary as caseload per year varied, we ran a linear regression. Consistent with Firth et al, (2015), we found that during years in which the therapist conducted more therapy sessions, her patients' outcomes were better (β = -0.004, p<0.001; see Table 3), although this effect was very small. Caseload accounted for 9.5% of the variance in patient outcomes. We did not find a significant effect of caseload on rate of collaborative terminations (β <0.001, p=0.32).

Hypothesis: Amount and magnitude of life events and professional responsibilities in the therapist's life would negatively predict patients' concurrent outcomes. To test the hypothesis that the therapist's patient outcomes would be worse when the therapist was experiencing stressful, exciting, and/or time-consuming personal and professional life events and responsibilities, we ran three separate regressions: each with the number of high stress, excitement, or time-consuming events in any given year predicting the adjusted, predicted end BDI score in that year. To account for the fact that during times in which the therapist had more going on in her personal and professional life she may have reduced her caseload, we investigated whether there was a correlation between number of events and concurrent caseload; there was a negative correlation between the two (r= -0.27, p< 0.001). Because of this, we accounted for concurrent caseload in each model. Consistent with our hypothesis, number of high excitement events predicted worse concurrent patient outcomes, even after accounting for concurrent caseload (β = 0.37, p< 0.001; see Table 3). The pattern was similar for number of high stress events predicting worse concurrent patient outcomes, but it was not significant (β = 0.17, p=0.06; see Table 2). Number of events that were time-consuming also did not significantly predict patient outcomes (β = -0.04, p = 0.27; see Table 3). Number of stressful,

exciting, or time-consuming events did not predict rates of collaborative terminations (p=0.14; p=0.31; p=0.27, respectively).

We ran post-hoc analyses to see whether personal or professional events were driving the results regarding high excitement events. Analyses accounted for concurrent caseload. Number of work events and obligations predicted worse concurrent mean end BDI scores, but number of personal events and obligations did not (β = 0.30, p<0.001; β = 0.23, p=0.15, respectively).

Hypothesis: The relationship between life events and professional responsibilities on patient outcomes would lessen as the therapist gained experience. To assess whether the therapist became better able to handle her life events and responsibilities, we created models that included mean-centered year, mean-centered number of one type of event/responsibility in that year, and an interaction between mean-centered year and number of the type of event. There was a significant interaction between number of high-stress events and year, such that an above-average number of high-stress events predicted significantly worse end BDI scores at the beginning of the therapist's career than did a below-average number of high-stress events, but, toward the end of the dataset, above-average and below-average numbers of high-stress events were predicting approximately the same adjusted end score (β = -0.03, p=0.02; see Figure 2). No other models were significant.

Study 1a Discussion

Results from Study 1a provided mixed evidence for the overarching hypothesis that the therapist would achieve better outcomes over time. As predicted, the therapist, on average, improved over the first 27 years of her career, as measured by average end BDI score, rate of collaborative termination, and plotting outcomes. She performed best during years in which she was more engaged in her clinical work (had a higher caseload) and performed worst in years in

which she was experiencing other stressful, exciting, or engaging life events or obligations—particularly professional ones. She became better equipped to utilize plotting outcome data with time, and plotting outcomes increasingly predicted collaborative terminations and lower end BDI scores as the years progressed. Occurrence of stressful events also predicted patients' end BDI score less over time, suggesting a potential resilience factor in her clinical work. While these results may indicate an effect of experience on effectiveness, it is also worth considering the possibility that other factors, such as continuing education and intentional learning, may have led to changes over time. Such factors were not measured in this study.

Some analyses did not yield significant results, however. Recording treatment goals and writing individualized case formulations did not increasingly predict lower end BDI scores over time, as predicted, and the therapist did not become more efficient in her work, as measured by her average predicted slope of improvement, as she gained experience. When we measured therapist improvement over time using individual patients' adjusted end BDI scores, rather than the average adjusted end score per year, improvement was not significant, though our primary analysis used average scores rather than individual scores. This difference in significance can be explained by the fact that, by only looking at the mean score per year, we are just looking at portion of variance related to mean levels, not individual levels.

Study 1b

Study 1a only accounted for the first three decades of this therapist's career; it is possible either that she: (a) continued to improve afterwards or (b) plateaued at a certain point, indicating a ceiling effect. The goals of Study 1b were to extend the results of Study 1a with a dataset that contains later years of the therapist's career. Our expectation was that the effects of time on therapist effectiveness would be smaller—or negligible—at this stage because enough years had

passed in the therapist's career that fewer gains could be made (given the likelihood of a ceiling effect). The study will utilize data from the DASS depression and stress scale, rather than from the BDI, as the therapist switched from the BDI to the DASS as her primary outcome measure for monitoring symptoms at each session of psychotherapy.

Method

Treatment

Treatment follows what was specified in Study 1a.

Procedure

Study 1b will utilize data from the DASS depression scale. Procedure follows what was specified in Study 1a.

Participants

This study will utilize data from the therapist who was the sole subject of Study 1a. Data were drawn from a database comprised of information from 150 adult patients who received individual therapy sessions during the years 2008-2021. For the purposes of Study 1b, we will use a subset of 129 adult patients who received individual therapy sessions during the years 2008-2019 by the same therapist. All patients gave written permission for the use of their treatment data for research purposes.

Patients were included in the present study if they had a Depression subscale score of 10 or higher or a Stress subscale score or 15 or higher on the DASS during the first and/or second therapy session; these are the cutoff for "mild" ranges of symptoms on the DASS (Lovibond & Lovibond, 1996). The sample that met these inclusion criteria consisted of 49 out of 129 (37.98%) patients who met criteria using the Depression cut-off score and 56 out of 129 (43.41%) patients who met criteria using the Stress cut-off score. When combining the two

subsets, there was a total of 62 out of 129 (48.06%) patients. This more inclusive sample was used for all analyses that used measures other than DASS scales as dependent variables. No demographic information was available for the sample.

Measures

The DASS is a measure with three subscales: depression, anxiety, and stress. Each subscale has seven items. Both the Depression and the Stress subscales were used to increase the comparability of Study 1a and 1b, given that items on the BDI are consistent with items on both the DASS depression and stress subscales.

Depression. We assessed symptoms of depression with the Depression scale of the DASS (Lovibond & Lovibond, 1995). The DASS is a widely-used, 21-item self-report measure of the severity of negative symptoms in the past week. The Depression scale contains 7 items and has been shown to have good internal consistency ($\alpha = 0.96$ for psychiatric patients; Brown, et al., 1997). The DASS Depression Scale and the BDI have been found to have a correlation of 0.74 (Lovibond & Lovibond, 1995). Example items from this subscale include "I couldn't seem to experience any positive feeling at all." and "I found it difficult to work up the initiative to do things."

Stress. We assessed symptoms of depression with the Stress scale of the DASS (P. F. Lovibond & Lovibond, 1995). The Stress scale contains 7 items and has been shown to have good internal consistency ($\alpha = 0.93$ for psychiatric patients; Brown et al., 1997). The DASS Stress Scale and the BDI have been found to have a correlation of 0.60 (P. F. Lovibond & Lovibond, 1995). Example items from this subscale include "I found it hard to wind down." and "I tended to over-react to situations."

Data Analysis Plan

Statistical procedures followed the models presented in Study 1a. Therapist experience was computed based on the number of years since 2008—the year she began administering the DASS; 2008 was defined as year 0. Analyses in Study 1b used individual predicted, adjusted end scores, rather than *mean* predicted, adjusted end score per year, as n-values in this dataset were small, and we wanted to prevent an outlier score from skewing the mean for that year.

For analyses that included variables that were also used in Study 1a, secondary analyses combined Study 1a and Study 1b data to look at change across 40 years, with the first year post-licensure as year 0.

Results

Therapist Experience as a Predictor of Patient Outcomes

Hypothesis: The therapist's outcomes, as measured by patients' symptom reduction, will be stable throughout Study 1b. To test the hypothesis that therapists' outcomes, as measured by symptom reduction, would no longer improve after nearly 3 decades, we ran two linear regressions in which years of experience predicted patients' standardized DASS depression or DASS stress end score. As predicted, years of experience did not significantly predict adjusted end DASS Depression score (β = 0.05, p= 0.85, 95% CI [-0.45, 0.55]) or DASS Stress score (β = 0.14, p= 0.54, 95% CI [-0.60, 0.32]; see Table 1).

Hypothesis: The therapist would maintain the lowered rate of uncollaborative terminations that she had achieved by the end of Study 1a. To assess whether the therapist maintained the relatively low rate of uncollaborative terminations that she had achieved by the end of Study 1, we conducted a logistic regression in which years of experience predicted rate of collaborative terminations. At baseline (in 2008), there was an 88.08% probability of the therapist's yielding a collaborative termination. It was found that the odds of yielding an

uncollaborative termination decreased by 8.01% (95% CI [-0.26, 0.12]) for each additional year of experience, although this finding was not significant (p= 0.41 see Table 2 for β). In post-hoc analyses, we investigated the change in collaborative terminations across Study 1 and 2 combined. It was found that the odds of yielding an uncollaborative termination decreased by 5.09% (95% CI [0.02, 0.08]) for each additional year of experience (β = 0.05, p<0.001).

Hypothesis: The therapist's use of decision-support tools (case formulation, plotting outcomes, and recording treatment goals) would continue at the level she had achieved by the end of Study 1a. To assess whether the therapist maintained the frequency of using decision-support tools that she achieved by the end of Study 1a, we conducted logistic regressions in which years of experience predicted rate of using each decision-support tool.

At baseline, her probability of plotting outcomes was 85.51%, of writing individualized case formulations was 39.99%, and of recording treatment goals was 68.84%. There was insufficient variability to calculate the odds of recording treatment goals for every additional year of experience (see Table 2 for β and p-value). The odds of her writing individualized case formulations and recording treatment goals, on the other hand, significantly increased by 19.31% (95% CI [0.03, 0.41]) and 39.39% (95% CI [0.06, 1.06]), respectively for each additional year of experience (see Table 2 for β and p-values).

When combining the Study 1a and b datasets, there was a significant quadratic effect for plotting outcomes and recording treatment goals (β = -0.0044, p<0.004; β = -0.001, p<0.001, respectively), indicating that the therapist increased in her use of these tools more at the beginning of her career and maintained a very high level of recording treatment goals after a certain point in her career. It was also found that the odds of writing individualized case

formulations decreased by 3.13% (95% CI [0.0093, 0.0099]) for each additional year of experience over her four-decade career.

Hypothesis: Greater caseload would predict better concurrent patient outcomes, as measured by standardized end scores. To test the hypothesis that the therapist's patient outcomes would improve as caseload per year increased, we ran two linear regressions, one for each of the DASS subscales. Caseload did not significantly predict concurrent DASS Depression or DASS Stress scores (β = 0.01, p=0.41; β = 0.01, p=0.56, respectively; see Table 3).

Hypothesis: Amount and magnitude of professional life events in the therapist's life will negatively predict patients' concurrent outcomes, but personal life events will not significantly predict outcomes. As with Study 1a, there was a negative correlation between number of events and concurrent caseload (r= -0.48, p< 0.001), so all models accounted for concurrent caseload. We ran four separate regressions: each with the number of professional or personal events in any given year predicting the adjusted, predicted end DASS depression or DASS stress score in that year. Contrary to our hypothesis, there was a significant negative effect of work events and obligations on both end DASS depression score (β = -1.85, p<0.001; see Table 3) and DASS stress score (β = -1.95, p<0.001; see Table 3), such that higher amounts of stressful events predicted better adjusted end scores. Also contrary to our hypothesis, number of personal life events had a significant linear effect on end DASS depression score (β = 1.09, p=0.002; see Table 3), such that higher amounts of personal events and obligations predicted worse adjusted depression scores. There was no significant effect of personal events on DASS stress score (β = 0.44, p=0.24).

Study 1b Discussion

The results of this study indicate that many elements of the therapist's evidence-based work were stable by the middle of the third decade of her career; there was no significant change over time in her patients' adjusted end depression or stress scores, the therapist's rate of collaborative terminations, or her use of plotting outcomes. Of course, we recognize the limits of overinterpreting a null hypothesis and finding. Other findings were not consistent with the expectation of no reliable change. The therapist's use of recording treatment goals and writing individualized case formulations did increase, despite predicting they would be stable. Also contrary to hypotheses, caseload did not significantly predict concurrent patient outcomes, as measured by stress scores and depression scores. Personal life events of the therapist predicted better concurrent patient outcomes, whereas professional life events predicted worse concurrent patient outcomes.

Study 1 Discussion

The combined results of Studies 1a and 1b indicate that the evidence-based therapist did improve in several ways as she gained experience. Therapist improvement was measured in terms of average patient symptom reduction, rate of collaborative terminations, use of decision-support tools, effectiveness in using decision-support tools, and ability to prevent life events from predicting concurrent therapy outcomes. The relationship between patient outcomes and caseload and personal and professional life events were also evaluated. In Study 1a, we found that there was, on average, improvement throughout the first 26 years of her career in some domains but not in others. In Study 1b, we found that the therapist no longer improved on some aspects of her practice, whereas, on some other dimensions (e.g., writing individualized case formulations and recording treatment goals), she continued to improve. It is unclear, however,

whether the differences in results across studies were due to the effects of time and experience or due to reduced power.

Not all hypotheses were supported, however. Plotting outcomes was the only decision-support tool in Study 1a that significantly interacted with time to predict lower adjusted end scores and higher rates of collaborative termination. Thus, it seems that the therapist's ability to use recording treatment goals and creating case formulation documents did not significantly improve over time.

Additionally, there was a significant quadratic effect of use of case formulations in Study 1a, indicating that she wrote written case formulations most frequently toward the middle of her career, and an overall negative linear effect of time on writing case formulations across Studies 1a and 1b combined. This is surprising, given that this therapist has spent much of her career pioneering the use of individualized case formulation in cognitive behavioral therapy. When compared to the other decision-support tools, however, writing individualized case formulations appears to be the most laborious of the three. To record treatment goals, the therapist uses a specific sheet and asks patients to complete the sheet for homework after the intake session. Similarly, a software tool now plots outcomes whenever she assigns a measure, which she generally always does. The case formulation, in contrast, is less formatted and also typically more time-consuming. It is also worth noting that the case formulation variable was only coded as present if there was a written formulation that was separate from any other clinical note (e.g., an intake). Thus, it is possible that the therapist could have begun integrating the formulation into other notes or creating the formulations in her head because she became confident in her ability to create formulations. It is also possible, however, that she may have been increasingly pulled away by other obligations, such as research; this trend would indicate drift.

Studies 1a and b also yielded some unexpected, contrasting results. Higher caseload significantly predicted better symptom outcomes in Study 1a but had no significant effects in Study 1b. One possible explanation for this may be that the therapist was more engaged in her clinical work at the beginning of her career and benefitted from having a high caseload in which she could gain informative feedback on her performance. By the third decade of her career, though, she was more focused on research and consultations, and her overall caseload was smaller than it had been. Because of this, her attention may have been divided, hindering her ability to learn from cases and apply the feedback to other cases. It is also worth noting that, on average, she conducted more therapy sessions per year in the first 27 years of her career than she did in the most recent 14 years of her career. It is also worth noting that the extent to which the therapist engaged in deliberate practice varied over time; for example, she consulted with other professionals to obtain informative feedback in some years but not in others. Thus, the relationships between deliberate practice and outcomes cannot be assessed fully.

Life events also indicated contrasting results between Studies 1a and 1b: Study 1a indicated that presence of other work events and obligations predicted worse patient outcomes, whereas Study 1b indicated that presence of other work events and obligations predicted better patient outcomes. These results may speak to a level of resilience that she achieved over time to prevent life events from negatively predicting psychotherapy outcomes. This would be consistent with Sherman and Thelen's (1998) finding that experience predicts less professional distress over time, although it also may be explained by the fact that her lower caseload in later years lessened her vulnerability to burnout.

These studies were the first to examine longitudinally whether an evidence-based therapist with decades worth of data improves over time. They also provide further evidence to

support Ericsson's (e.g., 2006, 2007) assortation that a detailed system of informative feedback, which has been indicated as necessary for improvement and expertise in other professions and hobbies, is likely important in psychotherapy. This idea may also explain why the therapist's increased use of specific decision-support tools significantly predicted better outcomes—both in terms of symptom reduction and rate of collaborative terminations: specifically, plotting outcomes and consistently reviewing them with the patient provides precise feedback on the course of therapy.

Study 2

The present study utilizes a group of 16 therapists—all trained and/or supervised by the psychologist on whom Study 1 was focused—in their first years of clinical work outside of graduate school training. Study 2 will build off of Study 1 by examining whether being trained by a therapist who engages in a system of deliberate practice results in improved performance over the relatively short training period.

Although prior research has indicated that trainees' education in how to engage in deliberate practice improves therapist-rated working alliance (Ferlie & Shortell, 2001), little—if any—research to date has tested for significant effects on therapy outcomes due to training from a deliberate practice framework. There is reason to believe that these therapists would benefit from being supervised by the director of the Center, as she has created a "culture for change" (Ferlie & Shortell, 2001) in which there is ongoing room for discussion about cases that are not progressing as expected (Goldberg, Babins-Wagner, et al., 2016). There is reason to believe that these trainees might improve at a faster rate than other therapists, given that their supervision provides another avenue for informed feedback on their performance and opportunity to adjust treatment as needed.

Hypotheses were preregistered at the Open Science Framework (https://osf.io/4qh3m). Based on the fact that the director showed progress in some domains over decades and that she designed a training program around methods with a sound theoretical relationship to learning and improvement over time, we predict that: (a) the therapists' patients would show greater symptom reduction in psychotherapy as the therapists gained experience; and (b) the therapists would experience fewer unilateral, uncollaborative terminations as they gained experience. Based on the results of Study 1, we also hypothesized that (c) amount and magnitude of life events and professional responsibilities in the therapists' lives will negatively predict patients' concurrent outcomes. Due to the differences in training history between director of the clinic and the therapists in this study, we also hypothesized that: (d) therapists' use of decision-support tools would increase in their first six months at the Center and then plateau, and (e) the therapists would become better able to use decision-support tools to yield fewer uncollaborative terminations and achieve more symptom reduction over time. Finally, because most therapists who participated in the life events interview listed high caseload as a stressor, we expected that (f) greater caseload would predict worse concurrent patient outcomes.

Method

This study includes 16 evidence-based therapists' data, which were collected in one sample that covers the years 1995-2006 and uses the Beck Depression Inventory (BDI) as the primary outcome measure. In the Spring of 2021, six of the therapists also participated in a video-recorded, semi-structured interview about their major personal and professional life events that occurred during the years in which they were at the Center; the level of stress, excitement/engagement, and time commitment for each event was also assessed.

Treatment

Treatment was the same as that detailed in Study 1a.

Participants

The data for this study were drawn from the archival database described previously. The present study invited the 19 therapists in the dataset who were not the founding director of the group practice to participate. Seventeen of the therapists agreed to participate; one of the 17 therapists was removed from the dataset, as he never received training or supervision from the founding director of the group practice. The 16 remaining therapists saw 674 patients over the course of the years data were collected for this dataset. Most therapists were Ph.D. psychologists; one was an M.S.W. All therapists were trained and/or supervised by the founding director, and their data in the database marks their first years after receiving their graduate degree.

Each of the 16 therapists was invited to participate in a semi-structured interview similar to the one described in Study 1; six of the therapists accepted the invitation. For each interview, the interviewer collected information about the therapist's level of experience (if they were in training still or if they were licensed) and about events in the therapist's personal and professional life over time.

Patients were included in the present study if they: (a) had BDI scores of 10 or higher during their first and/or second therapy session, (b) had not specified prior to attending a meeting that the purpose of meeting was consultation only, and (c) had more than one therapy session. A cutoff score of 10 was used, as this is the threshold for mild depression on the BDI. To simplify data analysis, when patients had more than one course of treatment within the dataset, we included only their first course.

The sample that met these criteria consisted of 207 of 674 (30.71%) patients. Information about participants' racial and ethnic identities is limited in that race and ethnicity were combined into a single question in the original dataset. Most of the sample (85.02%) identified as Caucasian, 4.35% as Asian, 4.35% as Hispanic, 1.93% as African American, and 2.42% as other. The remaining 1.93% did not have a recorded race on file. Most of the sample (57.97%) identified as female. Most patients (92.27%) were diagnosed with at least one anxiety or depressive disorder. Psychiatric diagnoses were made based on a clinical interview by the therapist, who used the most current version of the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1987, 1994, 2000) available at the time the patient was treated. The median first BDI score of the sample used was 19 (*SD* = 8.68). About half of patients (56.52%) were receiving adjunctive pharmacotherapy at the time of the intake and some (22.22%) received adjunctive psychosocial treatment (e.g., couples therapy, Alcoholics Anonymous).

Measures

All measures were the same as those detailed in Study 1a.

Data Analysis

We ran linear mixed model analyses in R (version 1.2.1335) using the lme4 package (Bates et al., 2014). Experience was measured in six-month intervals, with the first six months of each therapist's post-graduate training as time zero. Each interval included any therapy with an intake in that time period. Analyses measured improvement of up to the first two years of data, as this was a typical length for a post-doctoral fellowship at the Center. Statistical procedures to find adjusted, predicted end BDI scores replicated those described in Study 1a. For all analyses except for the ones noted in the results section, we used mean end score, as this

best allowed us to examine change in the therapist's average effectiveness over time, rather than individual patients' performance (as described in Study 1a); the sample size was big enough that, unlike Study 1b, we were able to do this. All models included a random intercept for therapists and a random slope to account for the fact that therapists likely vary in how rapidly they change or improve in their skill level. Only 10 of the 16 therapists had initial sessions that met inclusion criteria and started in more than one six-month interval; thus, the remaining six did not contribute to the slope of any analyses. Six of the therapists had four time points, two had three time points, and two had two time points.

Results

Preliminary Analyses

To account for the fact that patient diagnostic profiles might predict degree of treatment, three measures of symptom severity were considered in subsequent linear regressions: presence of a depression disorder diagnosis at intake, presence of an Axis II disorder diagnosis at intake, and total number of psychiatric diagnoses at intake. When each of the three measures was entered simultaneously into a linear regression model predicting patients' adjusted predicted end BDI scores, none of them was significant; thus, all were excluded from future regression models. For linear regressions involving interaction terms, we mean-centered predictor variables.

Therapist Experience as a Predictor of Patient Outcomes

Hypothesis: The therapists' patients would show greater symptom reduction in psychotherapy as they gained experience. To test the hypothesis that the therapists' outcomes, as measured by symptom reduction, improve over time, we created a linear mixed model growth curve with six-month intervals of time nested within therapists and experience predicting the mean adjusted end BDI score for each year. Specifically, the model included six-month intervals

of time (experience) as a fixed effect. It also included random effects of intercept and slope, to allow for the possibility that some therapists may have started as more effective than others or may have improved at faster rates than others. With only four time points and 12 therapists, the model was too complex to converge, so we excluded the random slope. When running the model without the random slope, there was a significant fixed effect of intercept (β = 13.08, SE=0.19, p<0.001), indicating that the average adjusted end BDI score during therapists' first six months was around 13. We did not find a significant main effect for therapist experience on patients' adjusted end BDI score (β =0.08, SE=0.19, p=0.67), suggesting that the therapists did not significantly improve over time (see Table 1 for variance).

Hypothesis: The therapists would experience fewer unilateral, uncollaborative terminations as they gained experience. We conducted a nonlinear mixed model growth curve with six-month intervals of time nested within therapists and experience predicting the rate of collaborative terminations for each year. The model included random effects of intercept and slope, to allow for the possibility that some therapists may have started as more effective than others or may have improved at faster rates than others. There was no significant fixed effect of therapists' experience on rate of collaboration (β = 0.16, SE=0.16, p= 0.31). There was a significant effect of the intercept (β >0.99, SE=0.24, p<0.001), indicating that therapists had a very high rate of collaborative terminations in their first six months at the Center (see Table 2 for variance). Premature termination was not considered, as therapists' leaving the Center during their last time interval was likely to confound the results.

Therapist Experience as a Predictor of Therapist Process

Hypothesis: Therapists' use of decision-support tools would increase in their first six months at the Center and then plateau. We conducted nonlinear mixed model growth curves

with six-month intervals of time nested within therapists and experience predicting therapists' frequency of using decision-support tools. We included a quadratic term, as we expected that therapists' uses of decision-support tools would increase when they first arrived at the Center and plateau thereafter. The models included random effects of intercept and slope. There was no significant fixed effect of therapists' experience on frequency of plotting outcomes (β = -0.32, SE=0.29, p= 0.26) or of the intercept (β = 0.92, SE=0.52, p= 0.08; see Table 2 for variance). There was no significant fixed effect of therapists' experience on frequency of writing individualized case formulations (β = -0.001, SE=0.03, p= 0.97). There was a significant fixed effect of the intercept (β = 0.66, SE=0.09, p<0.001; see Table 5 for variance), indicating that about two thirds of patients' records included a detailed, written case conceptualization in the first time interval. There was no significant fixed effect of therapists' experience on frequency of recording treatment goals (β = 0.01, SE=0.03, p= 0.56). There was a significant fixed effect of the intercept (β = 0.73, SE=0.08, p<0.001; see Table 5 for variance). Overall, these results do not support the hypothesis that therapists' use of decision-support tools would increase in the first six months and then plateau thereafter. They do indicate, however, that trainees adhered to these protocols with some consistency throughout their training at the Center.

Use of Decision-Support Tools as a Predictor of Patient Outcomes Over Time

Hypothesis: Therapists' experience would increasingly predict the relationship between use of decision-support tools and fewer uncollaborative terminations or more symptom reduction over time. To test the hypothesis that the therapists became better able to utilize decision-support tools over time to yield better results, we ran six mixed-model growth curves with six-month intervals nested within therapists: each model had one of the mean-centered decision-support tools, the mean-centered time variable, and an interaction between the

mean-centered tool and time as predictors and with either adjusted, predicted end BDI score or presence of a collaborative termination as an outcome. The models all included random effects of intercept and slope. There was a significant fixed effect of the interaction between time and plotting outcomes on adjusted end BDI score (β = -2.13, SE = 0.76, p = 0.01), indicating that experience predicted the relationship between plotting outcomes and adjusted end BDI score, such that therapists' patients had lower end BDI scores when their therapist plotted outcomes than they did when therapists did not plot outcomes. No other interactions were significant (see Tables 4 and 5).

Predictors of Patient Outcomes

Hypothesis: Greater caseload would predict worse concurrent patient outcomes. To test the hypothesis that therapists' caseload would predict patient outcomes, we created a linear mixed model growth curve with caseload predicting the mean adjusted end BDI score for each year. The model also included a random effect of intercept. Contrary to our hypothesis, there was no significant fixed effect of caseload (β = -0.001, SE = 0.002, p= 0.52). There was a significant fixed effect of intercept (β = 13.33, SE = 0.92, p<0.001; see Table 3 for variance).

Hypothesis: Amount and magnitude of life events and professional responsibilities in the therapists' lives will negatively predict patients' concurrent outcomes. To test the hypothesis that the therapists' patient outcomes would be worse when the therapists were experiencing personal and professional life events and responsibilities, we ran three mixed model growth curves with six-month intervals of time nested within therapists: each with the number of high stress, excitement, or time-consumption events as a fixed effect predicting the adjusted, predicted end BDI score in that year. The model also included a random effect of intercept. To account for the fact that during times in which therapist had more going on in their personal and

professional life they may have reduced their caseload, we investigated whether there was a correlation between number of events and concurrent caseload; there was no significant correlation between the two ($r = -0.09$, $p = 0.39$), so we did not account for concurrent caseload in the models. There were no significant effects indicating that stressful, exciting, or time-consuming events significantly predicted patients' concurrent outcomes (see Table 3).

Study 2 Discussion

The results of this study indicate that these evidence-based therapists did not improve in their first two post-grad years of their careers as measured by average patient symptom reduction, rate of collaborative terminations, and frequency of using decision-support tools. It is important to note, however, that power may be too limited by the small number of therapists involved in the study to observe anything but a very large effect. Out of the 16 therapists who consented to participate in the present study, only 12 contributed to the slope in each model, and only six of those had data for all four time intervals. Additionally, two years or less of deliberate practice may have been too brief to observe improvement. Despite this, plotting outcomes increasingly predicted lower adjusted end BDI scores as therapists gained experience. It is also possible, however, that the null findings might be true: improvement might not generally occur over time, even with the use of deliberate practice and decision-support tools. Future research could examine between-therapists differences to assess whether trends of change vary across therapists. Should this be the case, it may suggest different subgroups that benefit from time and experience in different ways, and hint at possible moderating effects that might distinguish the different subgroups. The ability to utilize this approach was limited in the current study due to the small number of therapists with data that spanned all four time points.

General Discussion

The present studies sought to address the question, *in what ways might evidence-based therapists get more effective as they gain "expertise?"* Broadly, the goals of these studies were to find whether time predicts change in patient outcomes (based on degree of symptoms, speed of symptom change, and type of termination) and therapist process (based on use of decision-support tools). We also sought to address the question of whether there are other factors that predict change in patient outcomes (e.g., therapist's caseload, therapist's life events).

Although the studies take an idiographic approach and cannot be generalized to larger populations, results from these studies exemplify the benefits of collecting session-by-session outcome data and utilizing it in a way that elicits immediate feedback on performance. Overall, the present studies suggest that it might be possible for evidence-based therapists who engage in clear systems of informative feedback to improve with experience but that further research is important to clarify this question. In Study 1a, the therapist did improve in the first few decades of her career. Study 1b, however, indicated that she stopped improving thereafter. In Study 2, there was, again, no evidence to suggest that the therapists improved with experience. These findings remained across patients with Axis II diagnoses and other indicators of symptom severity. Although these patient populations may be harder to treat, this finding is consistent with prior research that no patient factors, aside from initial symptom score, significantly predict magnitude of change in therapy (Hansen et al., 2002).

The studies also suggest that there are important ways, other than simply changes in symptom data, to measure therapist effects and whether therapists stay consistent across their careers in their therapy process (i.e., their use of clinical support tools and their ability to promote collaborative terminations). There is reason to believe, based on these studies, that evidence-based therapists should be mindful of their caseloads and should find a balance

between seeing enough patients that they are able to learn from feedback and apply to other patients but not seeing "too many." They also suggest that there may be important implications for therapists regarding reducing caseload or seeking consultation during times in which they have a lot of other events occurring in their lives. It is important to note that, in their respective life events inventories, therapists listed qualitatively different types of events; whereas some discussed national events such as the September 11 attacks, others solely considered events that occurred in their specific lives. It is possible, thus, that differences in results across studies might be, in part, explained by the impact of different types of events (including differential types and impact of events based on the therapists' particular stage of life).

In both Study 1 and Study 2, there was a significant interaction between time and plotting outcomes in predicting adjusted end BDI scores, such that plotting outcomes became a stronger predictor of lower BDI scores as the therapists gained experience; the same pattern was not observed for recording treatment goals and writing individualized case formulations. This may be an indicator of the importance of deliberate practice and clear systems of feedback, as plotting outcomes, more so than the other two decision-support tools, is a central contributor to systems of informative feedback that Ericsson (2006, 2007) detailed. Although the therapists in Study 2 did not significantly improve over time, it is worth noting that there were significant fixed effects of intercept for yielding collaborative terminations, recording treatment goals, and writing individualized case formulations. These results indicated that the therapists in Study 2 had very high (collaborative terminations and case formulations) and moderately high (treatment goals) rates of these outcomes in the first six months; this may indicate that the director of the Center (the therapist in Study 1a) effectively emphasized deliberate practice and decision-support tools from the start of therapists' time at the Center.

Despite these possibilities, the majority of analyses in Study 1b and Study 2 did not yield significant effects. Thus, the question of whether therapists who engage in deliberate practice improve over time remains one with mixed results, as do questions about the relationship between caseload and patient outcomes and the relationship between therapists' life events and patient outcomes. Further research is warranted to explore these questions further.

Limitations

Several limitations warrant note. Across all studies but particularly in Studies 1b and 2, power was likely very low due to a small number of observations. Therefore, results must be interpreted with caution. Future analyses should include a power analysis with these data to assess the likelihood of a type-one error. Further, the chance of a type-one error in Study 2 is particularly possible because of the low number of time points from which the data were measured. Future analyses could use days since a therapist's first therapy session, rather than six-month time intervals in Study 2 and years in Study 1, as a way to measure experience. Analyses could also use a therapist's total number of therapy sessions as a marker of experience, as time can overlook the varying number of patients that therapists see in a given day (Goldberg, Rousmaniere, et al., 2016). In Study 2, out of the 16 therapist participants, only six saw new patients who met exclusion criteria across all four six-month intervals. It is possible that those who stopped seeing new patients earlier had fewer opportunities to improve and, thus, could be skewing results. Future analyses with a larger sample could look solely at those who had two years' worth of intakes to see if there were significant effects for them. Analyses should also assess the possibility that patients who started in the last time interval of a therapist's tenure at the Center and had fewer than 20 sessions had shorter treatment durations and, thus, may have had less of an opportunity to improve. If this turns out to be the case, models should include a

level-two predictor to account for any cases in which the course of treatment ended because the therapist left the Center.

Studies 1a and 2 also utilized therapists' mean patient scores per year, rather than patients' individualized scores. Although the benefits of this method are discussed in the methods section of Study 1a, this approach also has its own limitations. Mean end scores did not calculate a weighted mean; years in which more patients met study criteria had the same weight as years in which fewer patients met criteria. Additionally, data are not missing at random. The therapists and patients may have been more likely not to continue with outcome monitoring when the patient is doing especially well or especially badly, for example. This can bias the data and may indicate that the data are not representative of the full sample. Follow-up analyses could consider using multiple imputation and/or Bayesian frameworks to account for the effects of missing data. The patient sample in these studies was also largely homogenous; the majority of patients were white and highly educated. Future research should explore these studies' questions with more heterogenous populations of therapists and patients, and in settings outside of private practice.

Finally, the life events interview was conducted retrospectively, requiring the therapist to recall events, as well as their stress and engagement levels, up to 40 years after they occurred. Prior research has found that the intensity of affect associated with an autobiographical memory decreases over time, particularly with negative memories (Walker et al., 2003); this may suggest that events were more stressful than the therapist recalled. Coding exciting and engaging events also limits our ability to understand specific associations between life events and patient outcomes; an event may be exciting but not engaging or vice versa, making it difficult to gauge

whether it is the positive valence of the events that predicts patient outcomes or the level in which the therapist was engaged in the event.

Despite these limitations, the current studies are important for highlighting the likely importance of using clear systems of informative feedback to gain expertise over time. Although the studies take an idiographic approach and cannot be generalized to larger populations, results suggest there are ways other than changes in symptom data to measure therapist effects and whether therapists stay consistent across their careers in their therapy process (i.e., their use of clinical support tools and their ability to promote collaborative terminations).

In summary, our results indicate that utilizing a clear system of informative, immediate feedback may be important for improving as a therapist over time, although further research is necessary. Although our results did not all consistently find improvement over time, they suggest a need for therapists to be aware of elements of the patients' improvement, their overall caseload, and other events in their personal and professional lives.

References

American Psychiatric Association. (1987). *Diagnostic and Statistical Manual of Mental Disorder (DSM-III-R)*.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM IV)*. Washington, DC: American Psychiatric Press.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (DSM-IV-TR)*. Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., & Grothendieck, G. (2019). *Package 'lme4.'*

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77–100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(56), 569.

Bickman, L. (1999). Practice makes perfect and other myths about mental health services. *American Psychologist*, *54*(11), 965.

Brown, T. A., Chorpita, B. F., Korotitsch, W., & Barlow, D. H. (1997). Psychometric properties of the Depression Anxiety Stress Scales (DASS) in clinical samples. *Behaviour Research and Therapy*, *35*(1), 79–89.

Budge, S. L., Owen, J. J., Kopta, S. M., Minami, T., Hanson, M. R., & Hirsch, G. (2013). Differences among trainees in client outcomes associated with the phase model of change.

Psychotherapy, *50*(2), 150.

Burns, L. R., & Wholey, D. R. (1991). The effects of patient, hospital, and physician characteristics on length of stay and mortality. *Medical Care*, 251–271.

Carlier, I. V. E., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J. A., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice*, *18*(1), 104–110.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., Thornton, J. A., & Andrews, W. P. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy*, *52*(3), 337.

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

Erekson, D. M., Lambert, M. J., & Eggett, D. L. (2015). The relationship between session frequency and psychotherapy outcome in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *83*(6), 1097.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge Handbook of Expertise and Expert Performance*, *38*, 685–705.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education*, *41*(12), 1124–1130.

Ferlie, E. B., & Shortell, S. M. (2001). Improving the quality of health care in the United

Kingdom and the United States: a framework for change. *The Milbank Quarterly*, *79*(2), 281–315.

Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research and Therapy*, *69*, 54–62.

Flood, A. B., Scott, W. R., Ewy, W., & Forrest Jr, W. H. (1982). Effectiveness in professional organizations: the impact of surgeons and surgical staff organizations on the quality of care in hospitals. *Health Services Research*, *17*(4), 341.

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

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology*, *63*(1), 1.

Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects and IAPT Psychological Wellbeing Practitioners (PWPs): A multilevel modelling and mixed methods analysis. *Behaviour Research and Therapy*, *63*, 43–54.

Gutner, C. A., Suvak, M. K., Sloan, D. M., & Resick, P. A. (2016). Does timing matter? Examining the impact of session timing on outcome. *Journal of Consulting and Clinical Psychology*, *84*(12), 1108.

Guy, J. D., Poelstra, P. L., & Stark, M. J. (1989). Personal distress and therapeutic effectiveness:

National survey of psychologists practicing psychotherapy. *Professional Psychology: Research and Practice*, *20*(1), 48.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*(2), 155–163.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*(3), 329.

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist–client feedback and clinical support tools. *Psychotherapy Research*, *17*(4), 379–392.

Hodges, N. J., Starkes, J. L., & MacMahon, C. (2006). Expert performance in sport: A cognitive perspective. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. J. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 471–488). Cambridge University Press.

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

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, *69*(2), 159.

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice

suggestions. *Journal of Clinical Psychology*, *61*(2), 165–174.

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

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, *10*(3), 288–301.

Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology & Psychotherapy*, *9*(2), 91–103.

Lehmann, A. C., & Gruber, H. (2006). Music. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. J. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 457–470). Cambridge University Press.

Leon, S. C., Martinovich, Z., Lutz, W., & Lyons, J. S. (2005). The effect of therapist experience on psychotherapy outcomes. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *12*(6), 417–426.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343.

Lovibond, S. H., & Lovibond, P. F. (1996). *Manual for the depression anxiety stress scales*. Psychology Foundation of Australia.

Luborsky, L., McLellan, A. T., Woody, G. E., O'Brien, C. P., & Auerbach, A. (1985). Therapist success and its determinants. *Archives of General Psychiatry*, *42*(6), 602–611.

Miller, S. D., Duncan, B. L., & Hubble, M. A. (2005). Outcome-informed clinical work. In *Handbook of psychotherapy integration* (Vol. 2). Oxford University Press New York, NY.

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

Nissen-Lie, H. A., Havik, O. E., Høglend, P. A., Monsen, J. T., & Rønnestad, M. H. (2013). The contribution of the quality of therapists' personal lives to the development of the working alliance. *Journal of Counseling Psychology*, *60*(4), 483.

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielsen, L., Dayton, D. D., & Vermeersch, D. A. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology*, *62*(9), 1157–1172.

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *10*(6), 361–373.

Orlinsky, D. E., & Rønnestad, M. H. (2005). *How psychotherapists develop: A study of therapeutic work and professional growth.*

Padesky, C. A., & Greenberger, D. (2012). *Clinician's guide to mind over mood*. Guilford Press.

Persons, J. B. (2012). *The case formulation approach to cognitive-behavior therapy*. Guilford Press.

Persons, J. B., & Hong, J. J. (2015). Case formulation and the outcome of cognitive behaviour therapy. In *Case formulation in cognitive behaviour therapy* (pp. 32–55). Routledge.

Persons, J. B., Roberts, N. A., Zalecki, C. A., & Brechwald, W. A. G. (2006). Naturalistic outcome of case formulation-driven cognitive-behavior therapy for anxious depressed outpatients. *Behaviour Research and Therapy*, *44*(7), 1041–1051.

Persons, J. B., & Tompkins, M. A. (1997). Cognitive-behavioral case formulation. *Handbook of*

Psychotherapy Case Formulation, 314–339.

RCore, T. (2016). *R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*.

Reneses, B., Munoz, E., & Lopez-Ibor, J. J. (2009). Factors predicting drop-out in community mental health centres. *World Psychiatry*, *8*(3), 173.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, *53*(2), 252–266.

Sherman, M. D., & Thelen, M. H. (1998). Distress and professional impairment among psychologists in clinical practice. *Professional Psychology: Research and Practice*, *29*(1), 79.

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

Shirk, S. R., & Karver, M. (2003). Prediction of treatment outcome from relationship variables in child and adolescent therapy: a meta-analytic review. *Journal of Consulting and Clinical Psychology*, *71*(3), 452.

Shirk, S. R., & Saiz, C. C. (1992). Clinical, empirical, and developmental perspectives on the therapeutic relationship in child psychotherapy. *Development and Psychopathology*, *4*(4), 713–728.

Skovholt, T. M., Rønnestad, M. H., & Jennings, L. (1997). Searching for expertise in counseling, psychotherapy, and professional psychology. *Educational Psychology Review*, *9*(4), 361–369.

Stein, D. M., & Lambert, M. J. (1995). Graduate training in psychotherapy: Are therapy outcomes enhanced? *Journal of Consulting and Clinical Psychology*, *63*(2), 182.

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547.

Tracey, T. J. (1986). Interactional correlates of premature termination. *Journal of Consulting and Clinical Psychology*, *54*(6), 784.

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.

Tryon, G. S., & Winograd, G. (2011). *Goal consensus and collaboration.*

Vocisano, C., Klein, D. N., Arnow, B., Rivera, C., Blalock, J. A., Rothbaum, B., Vivian, D., Markowitz, J. C., Kocsis, J. H., & Manber, R. (2004). Therapist Variables That Predict Symptom Change in Psychotherapy With Chronically Depressed Outpatients. *Psychotherapy: Theory, Research, Practice, Training*, *41*(3), 255.

Walker, W. R., Skowronski, J., Gibbons, J., Vogl, R., & Thompson, C. (2003). On the emotions that accompany autobiographical memories: Dysphoria disrupts the fading affect bias. *Cognition & Emotion*, *17*(5), 703–723.

Wampold, B. E. (2015). *Routine outcome monitoring: Coming of age—With the usual developmental challenges.*

Wampold, B. E., & Brown, G. S. J. (2005). Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology*, *73*(5), 914.

Westmacott, R., Hunsley, J., Best, M., Rumstein-McKean, O., & Schindler, D. (2010). Client and therapist views of contextual factors related to termination from psychotherapy: A

comparison between unilateral and mutual terminators. *Psychotherapy Research*, *20*(4), 423–435.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: the use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology*, *50*(1), 59.

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2017). Therapist effects on and predictors of non-consensual dropout in psychotherapy. *Clinical Psychology & Psychotherapy*, *24*(2), 312–321.

Table 1: *Therapist Experience as a Predictor of Patient Outcomes—Patient Scores*

Study	Effect
Study 1b Depression	**β=0.05**
	P=0.85
	R^2=0.02
Study 1b Stress	**β=-0.14**
	P=0.54
	R^2=0.01
Study 2	**Fixed effect β=0.08**
	Fixed effect SE=0.19

Table 2: *Therapist Experience as a Predictor of Patient Outcomes and Therapist Process*

	Study 1a	**Study 1b**	**Study 2**
Collaborative terminations	β=0.05 P=0.01 R^2=0.02	β=-0.08 P=0.41 R^2=0.01	*++Fixed effect β=0.16 Fixed effect SE=0.16 Variance=0.001
Premature terminations	β=0.004 P=0.85 R^2<0.001	N/A	N/A
Plotting outcomes	β=0.11 P<0.001 R^2=0.06	β=0.28 P=0.17 R^2=0.11	Fixed effect β=-0.33 Fixed effect SE=0.29 Variance=1.74
Written case formulations	*β=0.02 P<0.001	β=0.18 P=0.03 R^2=0.07	*++Fixed effect β<-0.01 Fixed effect SE=0.03 Variance=0.02
Written treatment goals	β=0.49 P<0.001 R^2=0.61	β=0.33 P=0.04 R^2=0.15	*++Fixed effect β=0.01 Fixed effect SE=0.03 Variance=0.01

*=quadratic model ++significant fixed intercepts (>0.99, 0.66, 0.73)

Figure 1

Study 1a: Use of plotting outcomes as a predictor of collaborative termination

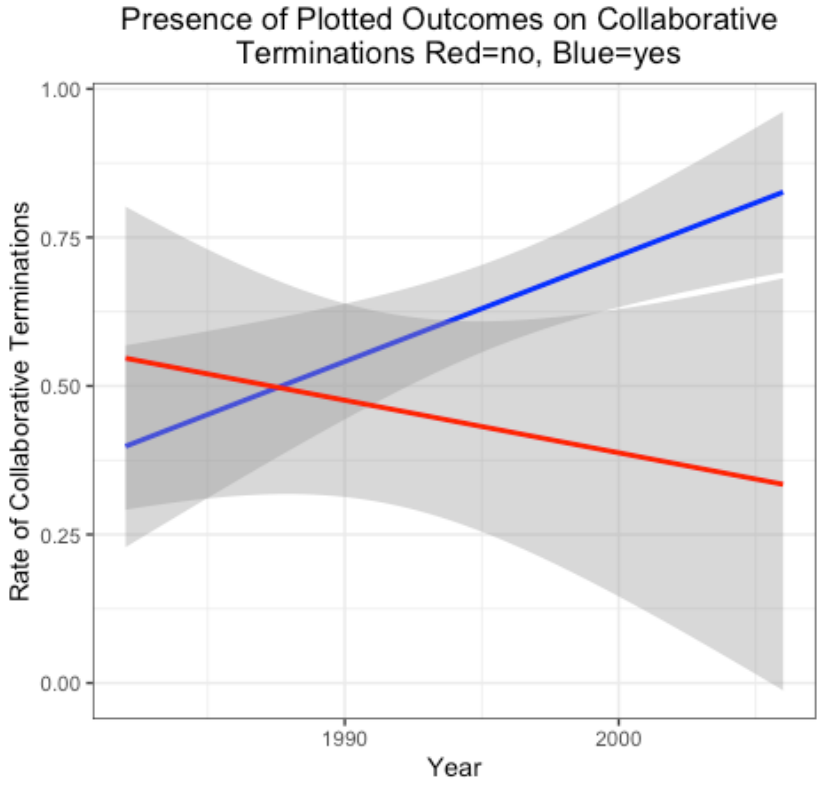

Table 3: *Personal and professional factors as predictors of patient outcomes*

Type of life event	**Study 1a**	**Study 1b Depression**	**Study 1b Stress**	**Study 2**
Caseload	**β=-0.004** **P<0.001** **R^2=0.09**	**β=0.01** **P=0.41** **R^2=0.01**	**β=0.01** **P=0.56** **R^2=0.01**	**Fixed effect β=-0.001** **Fixed effect SE=0.002**
High Stress	**β=0.16** **P=0.06** **R^2=0.11**	N/A	N/A	**Fixed effect β=-0.19** **Fixed effect SE=0.26**
High Excitement	**β=0.37** **P<0.001** **R^2=0.21**	N/A	N/A	**Fixed effect β=0.18** **Fixed effect SE=0.39**
High Time Consuming	**β=0.13** **P=0.18** **R^2=0.10**	N/A	N/A	**Fixed effect β=0.03** **Fixed effect SE=0.29**
Professional events	N/A	**β=-1.85** **P<0.001** **R^2=0.51**	**β=-1.95** **P<0.001** **R^2=0.44**	N/A
Personal events	N/A	**β=1.09** **P=0.002** **R^2=0.21**	**β=0.44** **P=0.24** **R^2=0.01**	N/A

Figure 2
Study 1a: Presence of high stress events on end BDI score

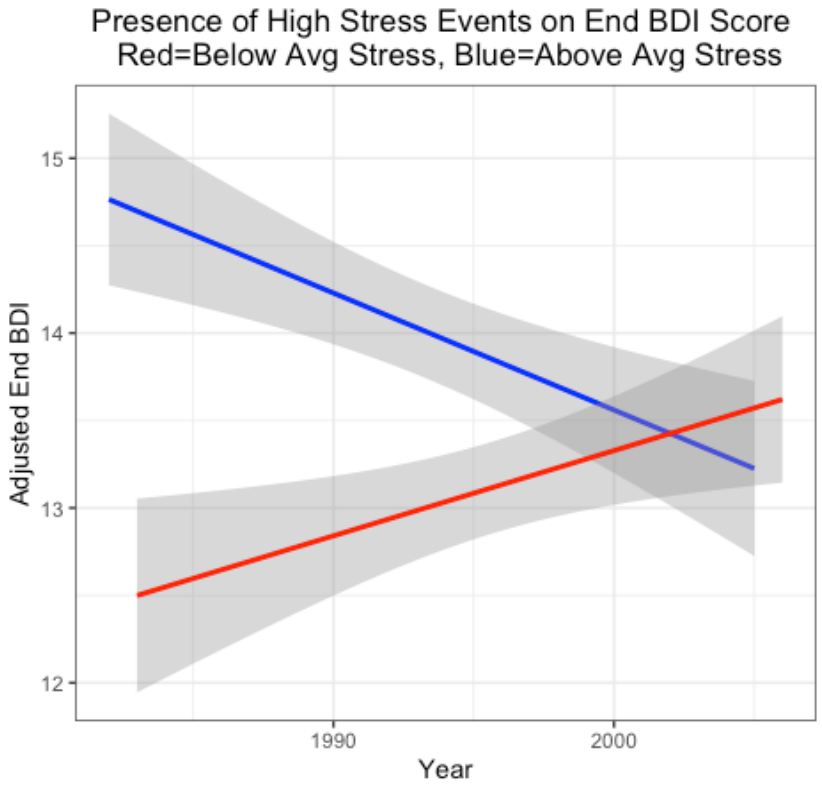

Table 4
Study 2: Experience predicting the relationship between use of decision-support tools and end BDI score, multiple therapists

	Fixed effect Beta	Fixed effect SE	Variance
Time*Plotting Outcomes			
Intercept	12.60	0.63	0.68
Experience	-2.13	0.76	0.05
Time*Case Formulations			
Intercept	13.42	0.65	0.43
Experience	-2.01	1.12	0.004
Time*Treatment Goals			
Intercept	13.29	0.62	0.16
Experience	-1.08	1.46	<0.001

Table 5
Study 2: Experience predicting the relationship between use of decision-support tools and collaborative terminations, multiple therapists

	Fixed effect Beta	Fixed effect SE	Variance
Time*Plotting Outcomes			
Intercept	0.78	0.04	<0.01
Experience	-0.08	0.06	0.0001
Time*Case Formulations			
Intercept	0.78	0.05	0.001
Experience	0.01	0.08	0.001
Time*Treatment Goals			
Intercept	0.80	0.04	0.004
Experience	-0.07	0.11	0.006