Psychotherapy Research, 2018
Vol. 28, No. 4, 532–544, https://doi.org/10.1080/10503307.2016.1216625

EMPIRICAL PAPER

Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists

SIMON B. GOLDBERG[1], WILLIAM T. HOYT[1], HELENE A. NISSEN-LIE[2], STEVAN LARS NIELSEN[3], & BRUCE E. WAMPOLD[1,4]

[1]*Department of Counseling Psychology, University of Wisconsin-Madison, Madison, WI, USA;* [2]*Department of Psychology, University of Oslo, Oslo, Norway;* [3]*Counseling and Psychological Services, Brigham Young University, Provo, UT, USA &* [4]*Modum Bad Psychiatric Center, Vikersund, Norway*

(Received 9 April 2016; revised 13 July 2016; accepted 14 July 2016)

Abstract

Objective: Differences between therapists in their average outcomes (i.e., therapist effects) have become a topic of increasing interest in psychotherapy research in the past decade. Relatively little work, however, has moved beyond identifying the presence of significant between-therapist variability in patient outcomes. The current study sought to examine the ways in which therapist effects emerge over the course of time in psychotherapy. **Method:** We used a large psychotherapy data set ($n = 5828$ patients seen by $n = 158$ therapists for 50,048 sessions of psychotherapy) and examined whether outcomes diverge for high-performing (HP) and low-performing (LP) therapists as treatment duration increases. **Results:** Therapists accounted for a small but significant proportion of variance in patient outcomes that was not explained by differences between therapists' caseload characteristics. The discrepancy in outcomes between HP and LP therapists increased as treatment duration increased (interaction coefficient = 0.071, $p < .001$). In addition, patients' trajectories of change were a function of their therapist's average outcome as well as the patient's duration of treatment (interaction coefficient = 0.060, $p = .040$). **Conclusions:** Indeed, patterns of change previously described ignoring between-therapist differences (e.g., dose-effect, good-enough level model) may vary systematically when disaggregated by therapist effect.

Keywords: therapist effects; trajectories of change; multilevel modeling; dose-effect; dose–response; good-enough level

Psychotherapy appears to be an effective method for improving patients' mental health (Seligman, 1995; Smith, Glass, & Miller, 1980). Patients, on average, tend to show considerable improvement in their psychological symptoms relative to those not receiving treatment (Wampold & Imel, 2015). While patient factors account for the lion's share of variability in treatment outcomes, therapists account for a significant proportion of variance as well (Baldwin & Imel, 2013). Recent meta-analytic estimates suggest that therapists' contributions to patient outcomes are on par with those of key therapeutic ingredients (e.g., therapeutic alliance; Horvath, Del Re, Flűckiger, & Symonds, 2011) and are considerably larger than differences between treatments (Wampold & Imel, 2015). On average, therapists account for between 3% and 7% of variability in patient outcomes (Baldwin & Imel, 2013).

Beyond establishing the presence of therapist effects in data drawn from a variety of contexts (e.g., naturalistic settings, randomized clinical trials), relatively little is known about how high-performing (HP) and low-performing (LP) therapists differ.[1] Some possibilities have been put forth in the literature, with HP therapists possessing higher degrees of facilitative interpersonal characteristics (e.g., warmth, empathy; Ackerman & Hilsenroth, 2003), more developed interpersonal skills (Schöttke, Fluckiger, Goldberg, Eversmann, & Lange, 2015), previous experience (Goldberg et al., 2016), or

Correspondence concerning this article should be addressed to Simon B. Goldberg, Department of Counseling Psychology, University of Wisconsin-Madison, 335 Education Building, 1000 Bascom Mall, Madison, WI 53706, USA. Email: sbgoldberg@wisc.edu

greater professional self-doubt (Nissen-Lie, Monsen, Ulleberg, & Rønnestad, 2013). Adherence to a treatment protocol, although theoretically an important therapist-level predictor of outcome, has not proven to be a strong predictor (Webb, DeRubeis, & Barber, 2010). Little is known about how actual outcomes differ for HP and LP therapists and how aspects of treatment, such as dosage, impacts outcomes.

Dosage is an important therapy variable, with significant economic and public health ramifications and relatively little consensus within the field regarding how long treatment should last. Since the beginning of psychotherapy, the field has argued about dosage. Freud recommended daily, hour-long meetings ("except on Sundays and public holidays," p. 367; Freud, 1989), with treatments commonly lasting years. Current trends have shifted towards treatments for particular disorders (Task Force on Promotion and Dissemination of Psychological Procedures, 1995) that are focused and of relatively brief duration, often between 8 and 12 weekly or biweekly sessions (e.g., Cognitive Processing Therapy for post-traumatic stress disorder, Resick, Nishith, Weaver, Astin, & Feuer, 2002; Interpersonal Psychotherapy for Depression, Cuijpers et al., 2011). Currently, few would recommend a lifetime's worth of psychotherapy, and some treatments report benefits from interventions as brief as a single session (Gingerich & Eisengart, 2000).

The impact of dosage may be a particularly important outcome by which to compare HP and LP therapists. If we imagine, for a moment, that HP and LP therapists represent two medications with differing degrees of effectiveness. One can likewise imagine differing dosage recommendations based on whether one was taking a strong, perhaps fast-acting HP medication, versus a less effective LP medication. Conveniently, treatment lengths vary widely in naturalistic data, providing ample opportunity to explore the ways in which HP and LP therapists may differ in the impact of dosage on their patients' outcomes.

Two primary models have been put forth in psychotherapy research to understand the impact of dose in psychotherapy: The dose-effect model and the good-enough level (GEL) model. Early work in the area of treatment dosage noted a dose-effect (or dose–response) of treatment duration, with longer courses of therapy generally associated with greater symptom reduction, albeit with some indication of diminishing returns (or negative acceleration) beyond a given dosage (Howard, Kopta, Krause, & Orlinsky, 1986). Later work proposed the "GEL" model as an explanation for this negative acceleration, suggesting that patients remain in therapy until they reach a GEL of symptom improvement (Barkham et al., 1996, 2006). From the GEL standpoint, patients (presumably in collaboration with their therapists) monitor their progress and make decisions about treatment length based in part on how their symptoms are responding to treatment. Of note, a symptom end point that is constant across durations (i.e., a GEL of symptomatology) allows the possibility that trajectories of change evidence a non-negatively accelerating dose-effect (e.g., a linear effect) occurring within a given length of treatment. This possibility could still match the negative acceleration noted by Howard et al. (1986) when examined in aggregate (see Barkham et al., 1996). In other words, it is possible that there is a dose–response in psychotherapy that does not slow down as treatment progresses (i.e., does not negatively accelerate as seen in the early dose–response literature; Howard et al., 1986), but that the overall rate of change varies systematically depending on the ultimate length of treatment (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009).

More recent work has supported the notion that trajectories indeed vary depending on ultimate length of treatment, with longer courses of therapy associated with flatter trajectories (Baldwin et al., 2009; Owen, Adelson, Budge, Kopta, & Reese, 2016; Stiles, Barkham, Connell, & Mellor-Clark, 2008; Stulz, Lutz, Kopta, Minami, & Saunders, 2013). Importantly, these longer courses of therapy are associated with symptom end points similar to shorter courses, supporting the notion that on average, patients reach a similar GEL near termination, regardless of treatment length.

Just as the GEL model sought to disaggregate change trajectories by length of treatment, it is of interest to examine the impact of treatment length disaggregated by therapists' overall effectiveness (i.e., by therapist effect). It could be that this is one of the key areas where differences between therapists are most pronounced—HP and LP therapists may differ starkly on the outcomes they are able to facilitate with their patients in short versus long courses of treatment.

This possibility—that HP and LP therapists differ in the impact of dosage (i.e., treatment length)—is a question of statistical interaction (i.e., moderation; Baron & Kenny, 1986). There are three primary theoretical possibilities worth noting here as to how therapists' overall effectiveness may impact the relationships between dosage and outcome. The most parsimonious possibility is that there is no interaction: The benefits of seeing an HP therapist instead of an LP therapist are uniform regardless of treatment length (i.e., parallel lines). An HP therapist produces superior outcomes for both short and long courses of therapy relative to the LP therapist. A second

possibility is that the difference in outcomes for HP versus LP therapists become more pronounced over time. Theoretically this could be due to both HP and LP therapists providing some (but not all) common factors (e.g., both are able to build basic rapport with patients). However, in the long run, it is the HP therapists who are able to structure longer courses of treatment and produce better results for cases that stay in treatment longer. The third possibility is that differences between HP and LP therapists are most pronounced initially and converge over time. In this instance, perhaps the HP therapists are able to engage patients quickly in treatment and facilitate change even in very brief therapeutic encounters. For patients who stay longer, however, the differences between HP and LP therapists become more closely linked to patient variables (e.g., motivation, insight, stage of change; Holdsworth, Bowen, Brown, & Howat, 2014). Regardless of which theoretical possibility may fit the data, demonstrating that therapists vary in the impact of dosage would have important implications for psychotherapy in naturalistic settings in which the length of treatment varies considerably.

The present study aimed to investigate to what extent patients' trajectories of change are influenced by systematic therapist-level differences (i.e., therapist effects). We sought to disaggregate outcomes by *both* treatment duration (as the GEL model proposes) *and* therapist, examining the possibility that trajectories of change vary systematically dependent on the therapist. In particular, we explored the emergence of therapist effects in naturalistic psychotherapy data (a setting in which dosage was not uniform across patients) across varying durations of treatment. We looked specifically at whether the trajectory of patient-level change is a function of both treatment duration and therapists' average outcome. Three research questions guided this work.

First, we hypothesized that therapist effects would be detected and not explained by caseload differences (e.g., caseload gender composition, baseline severity). This hypothesis was based on meta-analytic evidence supporting the robustness of therapist effects across various settings (Baldwin & Imel, 2013).

Second, we were interested in whether average patient outcomes would diverge across therapists as treatment duration increases. We did not have a directional hypothesis related to this research question but considered the three possibilities discussed above in which patients outcomes either converge, diverge, or are parallel across lengths of treatment.

Third, we were interested in whether trajectories of change differed depending on therapists' overall effectiveness (i.e., by therapist effect). Here again, we had no clear directional hypothesis. It was again theoretically plausible that trajectories were similar across HP and LP therapists with therapist effects accounted for by mean-level differences (rather than differences in trajectories of change) or that trajectories varied systematically dependent on therapist effect.

Method

Participants

Patients. Data were drawn from a large, naturalistic data set managed by the Research Consortium of Counseling and Psychological Services in Higher Education. These counseling centers provide a range of services including couples, group, and individual therapy to undergraduate and graduate students. As we were most interested in processes within individual therapy, the data set includes only patients who received individual therapy. The focal sample included 5828 patients seen by 158 therapists who were in the clinical range at baseline (i.e., Outcome Questionnaire-45 (OQ)-45 scores 63 or above; Lambert et al., 2004). Due to variation in data collection procedures, limited patient demographic data included only patients' age and gender. The sample included 3672 female patients (63.0%) and 2156 male patients. The average age was 22.63 ($SD = 4.11$, Mdn = 22.00, range = 16.89–59.65). Demographics reported previously regarding the counseling center consortium from which these data are drawn included the following ethnic/racial groups: 81.9% Caucasian, 3.4% Hispanic/Latino, 3.4% Asian/Asian American, 1.4% Indigenous American, 1.3% Pacific Islander, 0.8% African-American, 0.5% Other; 4.6% did not report.

Several data processing steps were necessary in order to arrive at this final sample and meet recommendations for employing multilevel modeling (MLM) with psychotherapy data (Baldwin & Imel, 2013). First, the sample was reduced to include only those patients who received at least three sessions of individual psychotherapy and completed OQ assessments with the same clinician. This was based on the rationale that fewer than three sessions would not adequately reflect a meaningful dose of treatment (Baldwin et al., 2009; Howard et al., 1986). Patients who saw multiple clinicians were excluded to avoid cross-classification between therapists. Further, we included only the first episode of care, defining a new episode when a patient either saw a new clinician or a period of 120 days or longer elapsed between sessions. Information was unavailable regarding whether final treatment sessions were planned terminations or not.

There was a definite positive skew in length of treatment in the current data, as has been noted in prior naturalistic data (e.g., Stiles et al., 2008). The mean number of sessions per patient in the sample was 8.59 ($SD = 8.47$, Mdn = 6, range = 3–153); a total of 50,048 sessions were included. Due to the notable skew in this distribution, winsorizing (Tukey, 1962) was used when examining treatment duration as a predictor in models. For these data, we employed a standard 5% cut-off, setting all data in the tails of the treatment duration distribution (5% on each side) to the duration value for the 5th and 95th percentile respectively. In practice, this only influenced the extreme high treatment durations, with all durations beyond the 95th percentile (longer than 22 sessions) set to the session length value for the 95th percentile.

Therapists. Psychotherapy was provided by 158 therapists, 65 female (41.1%) and 93 male. Description of these therapists' has been reported elsewhere (Goldberg et al., 2016, in press). Briefly, approximately 20.8% of therapists in the sample worked first as therapists in training (i.e., graduate students, predoctoral interns, or postdoctoral interns), then as licensed professionals. Approximately 30.5% of the sessions were provided by trainees, 38.7% provided by licensed professionals, and 30.8% provided by therapists who straddled these two statuses. As the focus of this work was on therapist effects, it was vital to assure that as reliable estimates of therapist effects as possible were obtained. To this end, the sample was reduced to include only patients whose therapists saw 10 or more patients within the clinical range data set. This data reduction step increases the reliability of therapist-level estimates (Baldwin, Imel, & Atkins, 2012; Crits-Christoph, Connolly Gibbons, Hamilton, Ring-Kurtz, & Gallop, 2011). Therapists saw on average 36.89 patients ($SD = 47.76$, Mdn = 18, range = 10–333).

The primary means to assign patients to therapist was based on available slots in the therapist schedules, although occasionally patients requested a therapist who was either a male or female and such requests were honored. Assignment was not based on patient severity, chronicity, or prognosis. Although assignment to therapist was not completely random, it could be described as quasi-random.

Measures

The outcome measure used in this data set was the OQ-45 (Lambert et al., 2004). This 45-item self-report measure was designed specifically to capture change that occurs during the course of psychotherapy. The measure has been widely used and shown to possess desirable psychometric properties, including high internal consistency reliability ($\alpha = 0.94$ for the total score in the current sample) and adequate test–retest reliability over a 3-week range (from 0.78 to 0.84; Snell, Mallinckrodt, Hill, & Lambert, 2001). Three subscales have been defined on the OQ-45: Symptom Distress (e.g., "I feel no interest in things," "I feel nervous"), Interpersonal Relations (e.g., "I am concerned about family troubles," "I have trouble getting along with friends and close acquaintances"), and Social Role Performance (e.g., "I feel that I am not doing well at work/school," "I feel stressed at work/school"). The use of the total score has been common practice and is supported by factor analytic work (Bludworth, Tracey, & Glidden-Tracey, 2010).

Statistical Methodology

Estimation of treatment effects. Standardized mean difference scores (i.e., Cohen's *d* [1988]) were computed at the patient-level using the difference between each patient's pre- and post-treatment OQ-45 total scores divided by the sample's pooled pre- and post-treatment standard deviation. These within-patient *d*s were included as the outcome in two-level models (patients nested within therapists) described below. As within-patient *d*s were computed as pre- minus post-treatment, a more positive effect size reflects a larger drop in symptoms during treatment.

Therapist-level outcomes were computed by taking the average of outcomes for all patients seen by a given therapist. This therapist aggregate *d* reflects differences in average outcome across therapists. A nice feature of computing therapist differences in this metric is the straightforward interpretation of therapists' aggregate effects as well as variation between therapists (e.g., the standard deviation of therapist aggregate *d*s provides a metric for assessing between-therapist variation complementary to the traditionally reported intraclass correlation coefficient [ICC]). The traditional ICC formula was also used (i.e., between-therapist variance divided by between- and within-therapist variance; Snijders & Bosker, 2012) with estimates derived from a two-level random intercept MLM (within-patient *d*s nested within therapists) with no additional predictors.

$$Y_{ij} = \beta_{00} + [U_{0j} + e_{ij}], \quad (1)$$

where Y_{ij} reflects the outcomes (within-patient *d*) of a given patient (*i*) seen by a given therapist (*j*). The

fixed intercept (reflecting the mean d across all therapists) was β_{00}, the random intercept (reflecting therapist-level deviation from the overall mean across therapists) was U_{0j}, and e_{ij} reflects the error or residual term.

Assessing potential caseload confounds. Several caseload characteristics were added as level 2 predictors to the initial two-level random intercept MLM (Equation (1)) in order to assess the degree to which between-therapist variation was explained by features of therapists' caseloads. Covariates were selected that could theoretically account for variation in patient outcomes (Goldberg et al., 2016). These included caseload average age, gender, baseline severity (i.e., baseline OQ score), treatment duration, number of cases terminating prior to session three, number of therapist's cases in the full sample, and number of therapist's cases in the clinical sample.

Two-level models: Modeling therapist differences across durations of treatment. Two-level MLMs (within-patient ds nested within therapists) were used to assess whether therapist overall effectiveness differentially impacted outcomes across varying durations of treatment. To address this question, a cross-level interaction term was included between therapist aggregate outcomes and patients' treatment duration.

$$\begin{aligned} Y_{ij} = {} & \beta_{00} + \beta_{10}(\text{Duration}) + \beta_{01}(\text{Therapist Aggregate } d) \\ & + \beta_{02}(\text{Duration} * \text{Therapist Aggregate } d) \\ & + [U_{0j} + e_{ij}], \end{aligned} \tag{2}$$

where Y_{ij} is the outcome (d) of a given patient (i) seen by a given therapist (j). This outcome was predicted by a fixed intercept (β_{00}), as well as by a given patient's treatment duration (β_{10}), a given therapist's aggregate d (β_{01}), the interaction between a given patient's treatment duration and a given therapist's aggregate d (β_{02}), along with a random intercept unique to each therapist (U_{0j}), and a residual term (e_{ij}).

A significant interaction between a patient's treatment duration and the therapist aggregate d would indicate that the gap between HP and LP therapists varies as a function of treatment duration. Subsequent models controlled for caseload characteristics as well.

Three-level models: Modeling therapist effects across trajectories of change. Three-level MLMs (session number nested within patient nested within therapist) were used to assess whether patients' trajectories of change varied by duration of treatment (as in the GEL model; Baldwin et al., 2009; Owen et al., 2016) as well as by therapist aggregate outcomes. To address this question, a three-way cross-level interaction term was included between therapist aggregate ds, patients' treatment duration, and session number. Subsequent models examined the addition of a random slope component, quadratic and cubic slope parameters as level-1 predictors (to allow the possibility that trajectories of change were non-linear, as has been suggested [Baldwin et al., 2009]), as well as lower level (i.e., two-way) interactions:

$$\begin{aligned} Y_{ijk} = {} & \beta_{000} + \beta_{100}\,(\text{Session Number}) \\ & + \beta_{200}\,(\text{Session Number})^2 \\ & + \beta_{300}\,(\text{Session Number})^3 \\ & + \beta_{010}\,(\text{Treatment Duration}) \\ & + \beta_{001}\,(\text{Therapist Aggregate } d) \\ & + \beta_{020}\,(\text{Session Number} * \text{Treatment Duration}) \\ & + \beta_{030}\,(\text{Session Number}^2 * \text{Treatment Duration}) \\ & + \beta_{040}\,(\text{Session Number}^3 * \text{Treatment Duration}) \\ & + \beta_{002}\,(\text{Session Number} * \text{Therapist Aggregate } d) \\ & + \beta_{003}\,(\text{Session Number}^2 * \text{Therapist Aggregate } d) \\ & + \beta_{004}\,(\text{Session Number}^3 * \text{Therapist Aggregate } d) \\ & + \beta_{005}\,(\text{Treatment Duration} * \text{Therapist Aggregate } d) \\ & + \beta_{006}\,(\text{Session Number} * \text{Treatment Duration} \\ & \quad * \text{Therapist Aggregate } d) \\ & + [U_{1ij} + U_{0jk} + U_{00k} + e_{ijk}]. \end{aligned} \tag{3}$$

In Equation (3), Y_{ijk} is the OQ-45 total score at a given session number (i) of a given patient (j) seen by a given therapist (k). This score was predicted by a fixed intercept (β_{000}), by linear (β_{100}), quadratic (β_{200}), and cubic (β_{300}) slope parameters as level-1 predictors. Treatment duration was entered as a patient-level (i.e., level 2) predictor (β_{010}). Therapist aggregate d was entered as a therapist-level (i.e., level 3) predictor (β_{001}). Two-way cross-level interactions were modeled between treatment duration and linear, quadratic, and cubic slope parameters (as in the GEL model; Baldwin et al., 2009), between therapist aggregate d and linear, quadratic, and cubic slope parameters, as well as between duration and therapist aggregate d (as in the previous model described in Equation (2)). The primary test of the third research question was provided in the three-way cross-level interaction between session number, duration, and therapist aggregate d. (Of note, additional three-way cross-level interactions were considered for the quadratic and cubic slope parameters but models failed to converge with these included.) Additional parameters included

random intercepts at the patient-level (U_{0jk}) and the therapist-level (U_{00k}) along with a residual term (e_{ijk}). A random slope component (U_{1jk}) also allowed patients' linear trajectories to vary around the overall linear effect. A significant three-way cross-level interaction would indicate that patients' trajectories of change are dependent on both treatment duration and therapists' overall outcome (i.e., aggregate *d*). As before, subsequent models controlled for patient and caseload characteristics as well.

Results

Descriptive Statistics

The initial overall mean within-patient *d* in the full sample ($n = 5828$) was $d = 0.99$ ($SD = 1.12$, Mdn = 0.88, range = −3.94 to 5.86). Using a three standard deviation cut-off, 34 patients were excluded ($n = 25$ [73.5%] high outliers, 9 [26.5%] low outliers), leaving a total of 5794 patients in the sample.[2] All therapists retained 10 or more cases in this restricted sample. Results reported from this point forward had outliers excluded.[3]

Descriptive statistics of patient- and therapist-level *d*s in the sample after exclusion of patient-level outliers are presented in Table I. Patients, on average, showed a large reduction in symptoms assessed on the OQ, with a mean within-patient $d = 0.98$ ($SD = 1.09$), with considerable variability in patient outcomes (*d*s from −2.35 to 4.32).[4] The mean therapist aggregate *d* was similar in magnitude ($d = 1.06$) and therapists as well showed variability in their average aggregate outcome ($SD = 0.33$, *d*s from −0.58 to 2.03), albeit much less so than the patient-level variability. Importantly, these effect sizes are on par with those achieved in both benchmarking studies and clinical trials of psychotherapy (cf., Minami, Wampold, Serlin, Kircher, & Brown, 2007).

Assessing the Validity of Therapist Effects

In order to examine the validity of therapist effect estimates in these data, random intercept MLMs were fit using within-patient *d*s as the outcome while specifying nesting within therapists. The ICC from the empty model (with no predictors) was 0.0089 (χ^2 [157] = 194.37, $p = .023$) indicating that slightly less than 1% of the variance in patient outcomes was explained at the therapist-level.[5] The ICC remained essentially unchanged with the seven caseload characteristics modeled, both individually and simultaneously (ICCs = 0.0086–0.0093). Thus the ICC appeared to be a function of the therapists themselves, rather than differences between caseloads.

Two-level Model Results: Therapist Differences Across Durations of Treatment

Additional two-level MLMs were next constructed in order to examine whether therapists' aggregate outcome impacted patient outcomes dependent on treatment duration. A two-level model was fit predicting within-patient *d*s from therapist's overall outcome (quantified as the therapist's aggregate *d*), the patient's length of treatment (as a level-1 predictor), and the interaction between therapists' aggregate *d* and length of treatment[6] (Table II). A subsequent model included the seven level 2 predictors.

The initial two-level model with no additional level 2 covariates showed a significant main effect for duration of treatment (winsorized), indicating that patients who remained in treatment longer showed poorer outcomes (estimate = −0.060, $p < .001$) when therapists' aggregate *d* and the interaction between length of treatment and therapists' aggregate *d* were held constant. As the primary test of the research question, a significant interaction was observed between therapists' aggregate *d* and length of treatment (estimate = 0.071, $p < .001$).[7] The direction of this effect can be interpreted as indicating that the association between-therapist effectiveness and patient outcome grows stronger as the length of treatment increases. In other words, differences in outcome for patients of effective and ineffective therapists increase with treatment length (Figure 1).

An additional model was constructed with the seven level 2 predictors added as fixed effect parameters. Examination of the fit indices (Bayesian Information Criterion [BIC] and Akaike Information Criterion [AIC]) suggested that these additional predictors did not improve the model fit beyond the simpler random intercept model (BIC and AICs increased with the inclusion of additional level 2 predictors: Initial model AIC = 17226.82, BIC = 17266.81; model with covariates AIC = 17281.98, BIC = 17368.59). Further, the interaction term of interest remained significant at $p < .001$ with the coefficient unchanged (estimate = 0.071) with these covariates included.[8] This suggests that the interaction between-therapist effect and length of treatment noted previously was not accounted for by differences in therapists' caseloads.

Three-level Model Results: Therapist Differences Across Trajectories of Change

Finally, three-level MLMs were constructed in order to examine whether patients' trajectories of change varied dependent on both treatment duration and therapists' overall outcome (i.e., aggregate *d*). As described below, formal model comparison was

Table I. Patient- and therapist-level descriptive statistics.

	n	Mean	*SD*	Median	Min	Max
Patient-level						
Within-patient *d*	5794	0.98	1.09	0.88	−2.35	4.32
% Female	5794	0.63	0.48	1.00	0.00	1.00
Age	5793	22.63	4.11	22	16.89	59.65
# Sessions	5794	8.57	8.41	6	3	153
# Sessions winsorized	5794	7.89	5.14	6	3	22
Baseline OQ total	5794	83.74	14.92	81	63	144
Therapist-level						
Therapist aggregate *d*	158	1.06	0.33	1.06	−0.58	2.03
% Female	158	0.64	0.14	0.63	0.33	1.00
Age	158	22.57	1.29	22.50	20.31	30.14
# Cases in full sample	158	102.97	132.32	49	17	821
# Cases in clinical sample	158	36.89	47.76	18	10	333
# Sessions	158	13.57	5.77	12.60	4.92	37.72
# Sesssions winsorized	158	7.94	1.52	7.84	4.38	12.50
Proportion staying 3+ sessions	158	0.87	0.08	0.88	0.61	1.00
Baseline OQ total	158	83.81	3.50	83.60	74.62	94.88

Notes. Proportion staying 3+ sessions computed on $n = 8416$ patients (i.e., estimates computed prior to excluding patients with fewer than 3 sessions). Pre-post *d* computed as pre-treatment minus post-treatment (i.e., positive values reflect a drop in symptoms). Descriptives computed once within-patient *d* outliers were excluded. OQ = Outcome Questionnaire-45 (Lambert et al., 2004); Min, minimum value; Max, maximum value.

conducted to arrive at a best-fitting model. The final model was a three-level model fit predicting session-level OQ total scores from linear, quadratic, and cubic slope parameters (as level 1 predictors), treatment duration (as a level 2 predictor), and therapists' aggregate *d* (as a level 3 predictor), and relevant two- and three-way interactions. The primary test of this hypothesis was provided in the three-way interaction between linear slope, treatment duration, and therapists' aggregate *d*. A subsequent model included the seven patient- and therapist-level (i.e., levels 2 and 3 in this model) predictors entered simultaneously.

Table II. Two-level multilevel model results predicting patients' pre–post change (within-patient *d*) from treatment duration and therapist overall outcome.

Predictor	Estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	0.78	0.10	5790	7.50	<.001
Treatment duration	−0.060	0.011	5790	−5.58	<.001
Therapist aggregate *d*	0.10	0.10	5790	1.02	.309
Treatment duration × therapist aggregate *d*	0.071	0.010	5790	6.90	<.001

Notes. Treatment duration interacts with therapists' overall outcome (i.e., aggregate *d*) to predict patient outcomes. Treatment duration represent winsorized treatment duration (due to positive skew in distribution of this variables). *SE*, standard error; *df*, degrees of freedom. $n = 5794$ patients, $n = 158$ therapists.

The initial three-level model tested included a single three-way interaction between linear slope, duration of treatment, and therapist's aggregate *d*. A second three-level model improved model fit by adding a random slope term that allowed linear slopes to vary (χ^2 [4] = 3548.40, $p < .001$). A third model added quadratic and cubic slope terms (with no additional interaction terms) also improving model fit (χ^2 [2] = 39.05, $p < .001$). A fourth model was attempted that included three-way interactions between quadratic and cubic slope terms with treatment duration, and therapist's aggregate *d* (along with the relevant two-way interactions). This model failed to converge. Thus a fifth model included the three-way interaction term between linear slope, treatment duration, and therapist's aggregate *d* along with additional two-way interactions between quadratic and cubic slopes with treatment duration as well as therapists' aggregate *d*. This more complex model significantly improved model fit (χ^2 [4] = 528.65, $p < .001$, Table II). A sixth model included the seven levels 2 and 3 control variables described. This model significantly improved fit (χ^2 [7] = 6326.40, $p < .001$) and was used as the final model (Table III).

As the primary test of the research question regarding whether trajectories of patient-level change depend both on therapists' average outcome (i.e., therapist's aggregate *d*) and treatment duration, a significant three-way interaction was detected in the final model (estimate = 0.060, $p = .040$). This indicates that patients' rates of change depend both on their duration of treatment as well as their therapists'

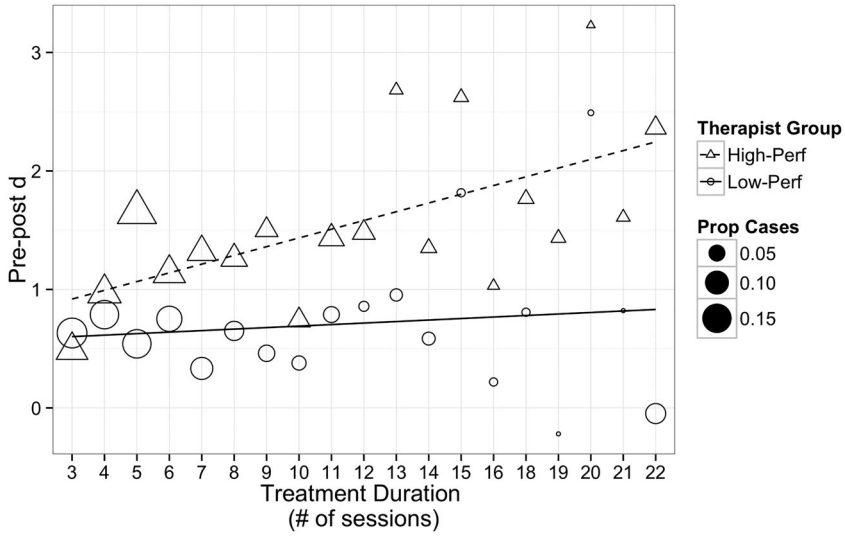

Figure 1. Therapists' aggregate *d* interacts with winsorized length of treatment to predict patient outcomes. Lines represent ordinary least squares best fit lines shown separately for the highest performing (High-Perf) 10% and lowest performing (Low-Perf) 10% of therapists in the data set ($n = 16$ therapists in each group). Therapists are separated in this way for plotting purposes only (i.e., models included therapist-level outcomes as continuous variables). Points in this figure represent the average outcome for patients seen by these therapists who received a given length of treatment. The size of points indicates the relative proportion (as %) of the therapists' cases that are represented by each value (i.e., Prop Cases in the figure legend).

average outcome. In addition, the two-way interactions between quadratic (and cubic) effects and duration imply a curvilinear trajectory of improvement, which varies as a function of treatment duration. Figure 2 displays model-derived trajectories for patients seeing either a HP (top 10%) or LP (bottom 10%) therapist for a given treatment duration. This figure displays the expected quadratic

Table III. Full three-level multilevel model results predicting patients' OQ total scores from linear, quadratic, and cubic time, treatment duration, and therapist overall outcome along with patient- and therapist-level covariates.

Predictor	Estimate	*SE*	*df*	*t*-value	*p*-value
(Intercept)	90.50	1.17	36,160	77.56	<.001
Session #	−7.57	0.45	36,160	−16.97	<.001
Treatment duration	−0.32	0.10	5630	−3.20	.001
Therapist aggregate *d*	3.16	1.08	152	2.94	.004
Session #2	0.90	0.05	36,160	17.58	<.001
Session #3	−0.035	0.0025	36,160	−13.67	<.001
Patient age	0.015	0.029	5630	0.51	.613
Patient gender	0.26	0.25	5630	1.02	.306
Patient baseline OQ	37.33	0.36	5630	103.30	<.001
Therapist average baseline OQ	6.53	2.28	152	2.86	.005
Therapist average treatment duration	−0.010	0.024	152	−0.40	.687
Therapist proportion staying 3+ sessions	5.28	2.19	152	2.41	.017
Therapist cases in clinical sample	−0.0017	0.0017	152	−1.02	.307
Session # × treatment duration	0.38	0.032	36,160	11.85	<.001
Session # × therapist aggregate *d*	−2.46	0.36	36,160	−6.90	<.001
Treatment duration × therapist aggregate *d*	−0.24	0.094	5630	−2.52	.012
Session #2 × treatment duration	−0.042	0.0023	36,160	−18.20	<.001
Session #2 × therapist aggregate *d*	0.028	0.0067	36,160	4.10	<.001
Session #3 × treatment duration	0.0016	0.00012	36,160	13.76	<.001
Session #3 × therapist aggregate *d*	−0.00022	0.000059	36,160	−3.78	<.001
Session # × treatment duration × therapist aggregate *d*	0.060	0.029	36,160	2.06	.040

Notes. Treatment duration and therapists' overall outcome (i.e., aggregate *d*) interact with time to predict patient OQ score. Treatment duration represents winsorized treatment duration (due to positive skew in distribution of this variables). Patient- and therapist-level covariates were grand mean centered. Session #, session number; *SE*, standard error; *df*, degrees of freedom. $n = 5793$ patients (demographics unavailable for $n = 1$ patient) seen by $n = 158$ therapists.

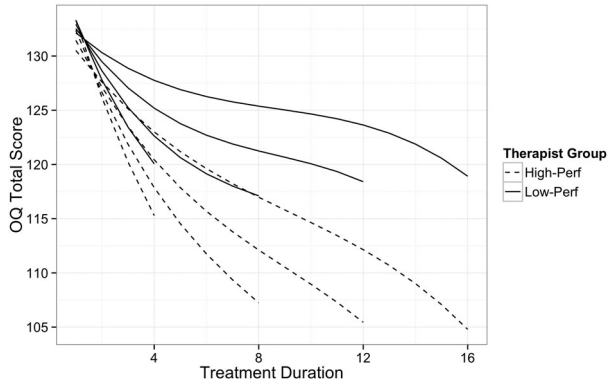

Figure 2. Therapists' aggregate *d* interacts with both treatment duration and linear slope. Lines represent model-derived trajectories for patients' seeing either an HP (top 10%) or LP (bottom 10%) therapist for a given treatment duration (i.e., 4, 8, 12, or 16 sessions). These extreme groups were selected for plotting purposes only—models included a continuous variable for therapists' aggregate *d*. Trajectories derived from full model that included seven patient- and therapist-level covariates (see Table III). Covariates were centered (to allow a meaningful intercept value) for plotting purposes. Note that the overall outcome trend (based on the low point of each of the curves) is in the opposite direction than in Figure 1 because lower scores on the OQ reflect improvement. OQ total = Outcome Questionnaire-45 total scores.

trend (i.e., diminishing marginal effect of sessions over time) that is qualified by a cubic trend, suggesting that after initial plateauing, outcome trajectories are characterized by a period of accelerated improvement as the end of treatment approaches. Note that the overall outcome trend (based on the low point of each of the curves) is in the opposite direction than in Figure 1 because lower scores on the OQ reflect improvement.

Discussion

The current study sought to move beyond merely estimating the magnitude of therapist effects in a large, naturalistic psychotherapy data set by examining the ways in which these differences emerge over the course of therapy. In particular, we were interested in whether patient outcomes diverged across time dependent on the therapist a patient saw.

As hypothesized, evidence was found suggesting that therapist effects were present and were robust to caseload confounds. It is worth considering, however, whether such a small proportion of variance being explained at the therapist-level—robust or not—is in fact a clinically meaningful finding. To address this question, it is important to put these effects into context. The current sample showed a substantial tendency to improve over the course of treatment ($d = 0.98$). This effect obscures, however striking variation between patients ($SD = 1.09$). Even with outliers excluded, patients' outcomes ranged from $d = -2.35$ to 4.32. A similar picture of variability emerges at the therapist-level. Overall, therapists tended to achieve a large reduction in symptoms across their caseload in aggregate (therapist-level aggregate $d = 1.06$). Therapists, like patients, also show significant variation in their average outcomes, albeit distributed more narrowly than at the patient-level. The highest performing therapist had a mean outcome roughly twice that of the average therapist ($d = 2.03$) while the lowest performing therapist had a mean outcome reflecting an increase in symptoms on average across his or her caseload ($d = -0.58$). Of course, given the ICC formula includes therapist- and patient-level variance in the denominator, a large amount of patient-level variance (relative to therapist-level variance) will necessarily produce a small ICC.

The small ICC (0.0089) may obscure the clinically meaningful ways in which patients' outcomes are likely to differ depending on whether one sees an HP or LP therapist.[9] The highest performing 10% of therapists (HP therapists) had a mean outcome three times that of the lowest performing 10% (LP therapists): ds = 1.61 versus 0.45 for HP and LP, respectively. This difference in average outcome (1.16) yields an number-needed-to-treat = 2, implying that only two patients would have to see an HP versus an LP therapist to achieve an additional clinical success.

With results supporting the existence, robustness, and clinical significance of variations between therapists in the current data, evidence was found suggesting that the gap between HP and LP therapists widens as the length of treatment increases. For patients who come for only three or four sessions, the outcomes achieved by HP and LP therapists are fairly similar. Based on the model-derived trajectories shown in Figure 2, patients attending four sessions of therapy and seeing an HP therapist could expect a drop of 18.01 on the OQ (reflecting $d = 1.00$) and those seeing an LP could expect a drop of 13.24 ($d = .73$). However, things look quite different at 16 sessions of therapy: A patient seeing an HP therapist could expect a drop of 25.71 on the OQ ($d = 1.42$) while a patient seeing an LP therapist could expect a drop of 13.22 ($d = .73$). Thus, it is not until one examines outcomes for longer courses of therapy that differences begin to emerge.

It seems reasonable to conclude, based on this finding that the quality of the therapist may be less important for short courses of treatment. There may be several potential explanations for this, varying on the type of treatment employed. (Of note, the following possibilities are purely theoretical as information at the level of specific techniques

employed was unavailable for this data set.) From a cognitive-behavioral therapy (CBT) standpoint, it may be that most therapists are able to provide basic psychoeducational frameworks and therapeutic advice in the initial sessions of therapy that are helpful for a subset of patients who stay only a few sessions. In longer courses of treatment, however, we could imagine the HP therapists are more effective in providing CBT-specific techniques: Identifying cognitive biases and maladaptive patterns in their patients (Beck, Rush, Shaw, & Emery, 1979; Young, 1999) and eliciting greater compliance with homework (Kazantzis, Whittington, & Dattilio, 2010).

From a common factors perspective (Wampold & Imel, 2015), it may be that most therapists are able to provide the basic facilitative conditions necessary for supporting change in very brief courses of therapy (i.e., three or four sessions) for relatively motivated patients with good prognoses. For those patients with more complex presentations, impoverished social support, and poorer prognoses (e.g., patients with personality disorders), however, differences between HP and LP therapists with greater interpersonal skill could have a profound impact on outcomes. Perhaps it is here that those therapists who are able to provide a more sophisticated rationale and a more nuanced ritual for treatment (Frank & Frank, 1991) as well as maintain an empathic stance in face of difficult interpersonal styles (e.g., interpersonal aggression) begin to demonstrate their therapeutic efficacy. Indeed, based on the plot displaying outcomes for the highest and lowest 10% of therapists, it would appear that the HP therapists have longer term cases marked by precisely this kind of progress. In contrast, the LP therapists tend to show a pattern of stagnation, with limited increased benefits for longer versus shorter courses of treatment.

Broadly speaking, the patients of HP therapists continued to receive benefits for staying in treatment for longer durations of care. Instead of showing a proposed GEL that is relatively constant regardless of duration, outcomes generally continued to improve as treatment length increased. This pattern appears to reflect more of a dose–response relationship, in which higher doses of treatment with an HP therapist is associated with greater therapeutic gains. In contrast, the LP therapists show a pattern that looks more like that predicted by the GEL model, in which a relatively uniform treatment effect is seen regardless of the dosage.

This possibility—that the dose-effect or the GEL model may apply differentially to HP and LP therapists—is an important one for unpacking both the therapist effect and trajectories of change. Just as change trajectories may vary depending on the length of treatment (e.g., in the GEL model; Baldwin et al., 2009), so the impact of treatment may vary depending on therapists' overall effectiveness.

Likewise, the impact of a therapist's overall effectiveness may vary depending on how long an individual stays in treatment. It appears from the current study that it is the long-term patients of the HP therapists who experience the greatest gains in treatment. For the short-term patients, it seems less important whether their therapist is HP or LP.

Clinically, these findings provide somewhat paradoxical recommendations for patients. Applying the medication analogy that is often invoked to discuss dosage in psychotherapy, one might assume that a higher dose would be required to receive benefits from a less effective therapist (akin to needing a higher dose of a weaker pain reliever for adequate relief). And, conversely, one might assume that fewer treatments with a highly effective therapist would be recommended. Our findings suggest the opposite, however. The observed pattern implies that if one is seeing an effective therapist, one may actually want to stay in treatment longer as one is likely to continue to benefit from therapy. In contrast, simply receiving a larger dose of treatment with a poorer performing therapist may not prove worthwhile.

Strengths, Limitations, and Future Directions

Strengths of the current study include a large sample of patients and therapists and the use of a standardized effect size (patient- and therapist-level *d*s). The use of standardized effect sizes was intended to facilitate interpretation of both patient and therapist outcomes in the current study and may be a worthwhile practice for future studies of therapist effects. In particular, the standard deviation of therapist-level outcomes may provide a useful complement to the traditionally reported ICC as a metric of between-therapist variation.

Limitations include a single self-report measure of outcome (OQ-45), limited patient- and therapist-level variables (e.g., no assessment of patient motivation, therapeutic alliance, or therapist factors) and a single setting (i.e., counseling center). Indeed, observed effects may not be expected to replicate in another setting in which session length, for example, was more tightly controlled (e.g., health maintenance organization setting). Substantively if indeed results do not replicate in a different setting, it may indicate that what it means to be an HP therapist is partially dependent on setting of practice. In addition, the student population included may also have impacted the degree of between-therapist

variability (i.e., reduced the ICC due to lower overall severity; Saxon & Barkham, 2012) or increased the within-patient variability. Another limitation was the lack of data regarding whether final sessions were planned. A data set with this information could be used to explore whether findings hold when only treatment completers are examined (i.e., when individuals who dropped out prematurely are excluded or otherwise modeled). Further, the current analyses were limited by requiring only 10 patients per therapist (Soldz, 2006). Our decision to require only 10 patients per therapist allowed the inclusion of a larger number of therapists in the sample (than requiring, for example, 20 or 30 patients per therapist). However, having only 10 patients may have reduced the reliability of therapist-level estimates of outcome.

It will be important in future studies to attempt to replicate the current relationships in another sample, including a non-counseling center sample. This is particularly important given the small therapist effect detected in the current sample, which is considerably lower than has been reported even for other university counseling center samples (e.g., 4.3% in Schiefele et al., 2016). Further work may also benefit from examining differences between HP and LP therapists themselves (e.g., assessing therapists' capacity for empathy using behavioral measures of empathy; Zaki, Bolger, & Ochsner, 2008). In addition, as treatment length has been associated with outcomes in important ways (in this study and elsewhere), it may be intriguing to assess the presence of therapist effects in treatment lengths (and early termination) and the implications of these therapist differences for patient outcomes. Ultimately, work will need to identify therapist characteristics most closely tied to patient outcomes and assess the possibility of supporting these qualities in trainees and therapists in practice (see Goldberg et al., in press). In addition, work seeking to understand trajectories of psychological change may do well to model therapist differences directly and examine whether trajectories vary systematically as a function of therapists' average outcome.

Acknowledgements

A previous version of this paper was presented at the 46th International Meeting of the Society for Psychotherapy Research (SPR), June 2015, Philadelphia, PA, USA.

Notes

[1] Of note, we refer here to dichotomous groups of HP and LP therapists and later to short versus long courses of treatment. Importantly, these groupings are referenced simply for rhetorical purposes. Both therapists' average effectiveness and treatment duration were treated continuously in all models.

[2] Univariate descriptive statistics were computed on patient-level outcomes (within-patient *d*) in order to assess the presence of outliers. As a method for addressing outliers, we employed a typical three standard deviation cut-off. Although one would expect some percentage of cases to be this deviant from the mean, given the large sample, it seemed worthwhile employing a standard cut-off that was likely to exclude cases that may be artifacts of estimation procedures and data entry errors. In order to assess the impact of outliers, primary analyses were run with and without these patients excluded.

[3] No differences were noted in results from primary analyses with or without outliers included.

[4] The distribution of within-patient *d*s was relatively normally distributed with some evidence of a positive skew based on inspection of a QQ-normal plot.

[5] The ICC was also computed in a model predicting post-test OQ scores controlling for pre-test OQ scores. The ICC was similar with this method (ICC = 0.011).

[6] An additional model was fit with a random slope term that allowed the relationship between therapists' aggregate *d* and within-patient *d* to vary across therapists. A χ^2 log-likelihood ratio test was used to assess improvement in model fit. The random slope model showed no indication of improved fit ($\chi^2 < 0.01$, $p = .999$) thus the random intercept model was used.

[7] An alternative method was also used to assess the impact of treatment length depending on therapists' overall outcome. This was done due to concerns of over-fitting the two-level MLM by including a predictor variable (therapists' average outcome) that was statistically composed of the outcome (within-patient *d*s). Specifically, we examined the correlation between the random slope and the random intercept in a model that included treatment duration as the only predictor of within-patient *d*s. We were interested in interpreting the slope-intercept correlation. This correlation can be interpreted as reflecting the relationship between therapists' average outcome (i.e., the random intercepts) and the model's random slopes (which reflect the relationship between duration and patient outcome). In this model, treatment duration was centered within therapist (so it could be interpreted as reflecting the intercept for treatment duration for each therapist's average patient, rather than the intercept for a treatment duration of zero). The slope–intercept correlation in this model was quite large, $r = 0.70$. This correlation is consistent with the two-way interaction reported between treatment duration and therapists' average outcome, with HP therapists (i.e., with higher intercepts) showing larger reductions in symptom over time (i.e., higher slopes). As the "nlme" package does not report a confidence interval for the slope–intercept correlation, we ran a series of bootstrapped replications ($n = 10{,}000$) with replacement. The empirical 95% CI from this was [−0.01, 0.67] which was a marginally significant effect. Of note, this test may be underpowered, specifically due to the need to center the treatment length variable within therapists. This means that the intercept reflects the therapist's average effect, but not very exactly. For brief-treatment dyads, the intercept reflects a forecast of client status at a session not actually measured. For long-term dyads, the intercept reflects an intermediate outcome sometime during treatment. As the therapist variance increased with treatment length in this data set, this approach gives too little weight to longer term treatment outcomes, so likely attenuates the correlation between therapists' average outcome and treatment duration. Based on comments from an anonymous reviewer, an additional model was run that included a quadratic term for length of treatment. The interaction noted previously between therapists' aggregate *d* and

length of treatment remained essentially unchanged (estimate = 0.070, $p < .001$) when the quadratic term was included. Further, in a model that added an interaction between this quadratic term and therapists' aggregate *d* found the interaction to be non-significant (estimate = 0.00033, $p = .858$).

[8] The interaction remained significant when the covariates were entered individually as well.

[9] It is also worth noting that a small ICC generally biases against the reported findings. A small amount of between-therapist variance relative to total variance reduces statistical power and may reflect a restricted range of between-therapist outcomes. Thus, one could interpret this small ICC as a conservative (rather than liberal) source of bias. The small ICC does not, however, necessarily bias against detecting the current findings. For example, an unusually low amount of between-therapist variance relative to total variance in the brief courses of treatment (e.g., three to four sessions) but not in the longer courses of treatment could influence the observed findings. Thus, it is important to replicate the current findings in another sample, ideally one with more typical therapist effects (e.g., Schiefele et al., 2016).

References

Ackerman, S. J., & Hilsenroth, M. J. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review*, *23*(1), 1–33.

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

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). Hoboken, NJ: Wiley & Sons.

Baldwin, S. A., Imel, Z. E., & Atkins, D. C. (2012). The influence of therapist variance on the dependability of therapists' alliance scores: A brief comment on "The dependability of alliance assessments: The alliance-outcome correlation is larger than you think" (Crits-Christoph et al., 2011). *Journal of Consulting and Clinical Psychology*, *80*(5), 947–951. doi:10.1037/a0027935

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

Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology*, *64*(5), 927–935.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York, NY: Guilford Press.

Bludworth, J. L., Tracey, T. J. G., & Glidden-Tracey, C. (2010). The bilevel structure of the outcome questionnaire – 45. *Psychological Assessment*, *22*(2), 350–355. doi:10.1037/a0019187

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crits-Christoph, P., Connolly Gibbons, M. B., Hamilton, J., Ring-Kurtz, S., & Gallop, R. (2011). The dependability of alliance assessments: The alliance–outcome correlation is larger than you might think. *Journal of Consulting and Clinical Psychology*, *79*(3), 267–278. doi:10.1037/a0023668

Cuijpers, P., Geraedts, A. S., van Oppen, P., Andersson, G., Markowitz, J. C., & van Straten, A. (2011). Interpersonal psychotherapy for depression: A meta-analysis. *American Journal of Psychiatry*, *168*(6), 581–592.

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy* (3rd ed.). Baltimore, MD: Johns Hopkins University Press.

Freud, S. (1989). On beginning the treatment. In P. Gay (Ed.), *The Freud reader* (pp. 363–377). New York, NY: Norton.

Gingerich, W. J., & Eisengart, S. (2000). Solution-focused brief therapy: A review of the outcome research. *Family Process*, *39*(4), 477–498.

Goldberg, S. B., Babins-Wagner, R., Rousmaniere, T., Berzins, S., Hoyt, W. T., Whipple, J. L., … Wampold, B. E. (in press). Creating a climate for therapist improvement: A case study of an agency focused on outcomes and deliberate practice. *Psychotherapy*. doi:10.1037/pst0000060

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., … Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology*, *63*(1), 1–11. doi:10.1037/cou0000131

Holdsworth, E., Bowen, E., Brown, S., & Howat, D. (2014). Client engagement in psychotherapeutic treatment and associations with client characteristics, therapist characteristics, and treatment factors. *Clinical Psychology Review*, *34*, 428–450. doi:10.1016/j.cpr.2014.06.004

Horvath, A. O., Del Re, A. C., Flűckiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*(1), 9–16. doi:10.1037/a0022186

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist*, *41*(2), 159–164.

Kazantzis, N., Whittington, C., & Dattilio, F. (2010). Meta-analysis of homework effects in cognitive and behavioral therapy: A replication and extension. *Clinical Psychology: Science and Practice*, *17*(2), 144–156.

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., … Burlingame, G. B. (2004). *Administration and scoring manual for the outcome questionnaire-45.* Orem, UT: American Professional Credentialing Services.

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology*, *75*(2), 232–243. doi:10.1037/0022-006X.75.2.232

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

Owen, J. J., Adelson, J., Budge, S., Kopta, S. M., & Reese, R. J. (2016). Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research*, *26*(1), 22–30. doi:10.1080/10503307.2014.966346

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology*, *70*(4), 867–879.

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology*, *80*(4), 535–546. doi:10.1037/a0028898

Schiefele, A. K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., … Lambert, M. J. (2016). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Administration and Policy in Mental Health and Mental Health Services Research*. Advance online publication. doi:10.1007/s10488-016-0736-3

Schöttke, H., Fluckiger, C., Goldberg, S. B., Eversmann, J., & Lange, J. (2015). Predicting psychotherapy outcome based on therapist interpersonal skills: A five-year longitudinal study of a therapist assessment protocol. *Psychotherapy Research*. doi:10.1080/10503307.2015.1125546

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The consumer reports study. *American Psychologist*, *50*(12), 965–974.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. London: Johns Hopkins University Press.

Snell, M. N., Mallinckrodt, B., Hill, R. D., & Lambert, M. J. (2001). Predicting counseling center clients' response to counseling: A 1-year follow-up. *Journal of Counseling Psychology*, *48* (4), 463–473. doi:10.1037//0022-0167.48.4.463

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

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research*, *16*, 173–177. doi:10.1080/10503300500264937

Stiles, W. B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008). Responsive regulation of treatment duration in routine practice in United Kingdom primary care settings: Replication in a larger sample. *Journal of Consulting and Clinical Psychology*, *76* (2), 298–305. doi:10.1037/0022-006X.76.2.298

Stulz, N., Lutz, W., Kopta, S. M., Minami, T., & Saunders, S. M. (2013). Dose-effect relationship in routine outpatient psychotherapy: Does treatment duration matter? *Journal of Counseling Psychology*, *60*(4), 593–600. doi:10.1037/a0033589

Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically-validated psychological treatments: Report and recommendations. *The Clinical Psychologist*, *48*(1), 2–23.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Wampold, B., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). New York, NY: Routledge.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78* (2), 200–211. doi:10.1037/a0018912

Young, J. E. (1999). *Cognitive therapy for personality disorders: A schema-focused approach* (Rev. ed.). Sarasota, FL: Professional Resource Press.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science*, *19* (4), 399–404.