

Therapist Perceptions of Their Own Measurement-Based, Problem-Specific Effectiveness

Michael J. Constantino¹, James F. Boswell², Alice E. Coyne³, Heather J. Muir¹,
Averi N. Gaines¹, and David R. Kraus⁴

¹ Department of Psychological and Brain Sciences, University of Massachusetts Amherst

² Department of Psychology, University at Albany, State University of New York

³ Department of Psychological Sciences, Case Western Reserve University

⁴ Outcome Referrals Inc., Framingham, Massachusetts, United States

Objective: Patient-reported outcomes data reveal differences both in therapists' global effectiveness across their average patient (between-therapist effect) and in treating different problems within their caseload (within-therapist effects). Yet, it is unclear how accurately therapists perceive their own measurement-based, problem-specific effectiveness and whether such self-perceptions predict global between-therapist performance differences. We explored these questions in naturalistic psychotherapy. **Method:** For 50 therapists, we drew on data from a mean of 27 past patients (total $N = 1,363$) who completed a multidimensional outcome measure—Treatment Outcome Package (TOP)—at pre- and posttreatment. For each of 12 outcome domains (e.g., depression, anxiety), TOP data classified therapists as historically “effective,” “neutral,” or “ineffective.” Unaware of their data-driven classifications, therapists rated their perceived effectiveness for each domain. We conducted chi-square analyses to determine whether therapists predicted their own measurement-based effectiveness classifications to a level greater than chance. We then used multilevel modeling to test whether therapists' problem-specific perceptions predicted global between-therapist performance differences. **Results:** For all but one outcome domain, therapists were no better than chance at predicting their measurement-based effectiveness classification. Additionally, controlling for patient baseline impairment, therapists who consistently overestimated their problem-specific effectiveness had patients who reported worse global outcomes than patients whose therapist more accurately estimated their effectiveness. Conversely, therapists who underestimated their problem-specific effectiveness had patients who reported better outcomes than patients whose therapist over- or accurately estimated their effectiveness. **Conclusions:** Therapist humility may differentiate the most from least globally effective therapists, and this virtue should be cultivated in clinical trainings.


What is the public health significance of this article?


Psychotherapy patients often choose a therapist based on the therapist's self-assessed areas of expertise. Yet, this study indicated that therapists can be inaccurate judges of their own measurement-based, problem-specific effectiveness, which underscores a significant public health issue. Moreover, the direction of such inaccuracies may be clinically meaningful; whereas therapist underestimation of their measurement-based effectiveness classifications (humility) predicted better between-therapist (patient-reported) global outcomes in naturalistic treatment, therapist overestimation (overconfidence bias) predicted worse between-therapist global outcomes. These results reinforce the importance of using an outcome measure to determine, and transparently report to the public, therapists' problem-specific effectiveness.


Keywords: therapist self-perceived effectiveness, within-therapist effectiveness differences, between-therapist effectiveness differences, therapist humility, naturalistic psychotherapy

Supplemental materials: <https://doi.org/10.1037/ccp0000813.supp>

Michael J. Constantino  <https://orcid.org/0000-0003-3126-2575>

James F. Boswell  <https://orcid.org/0000-0001-6214-0787>

Alice E. Coyne  <https://orcid.org/0000-0002-5950-0486>

Averi N. Gaines  <https://orcid.org/0000-0001-5856-7059>

Research reported in this article was supported by a Patient-Centered Outcomes Research Institute (PCORI) Award (IHS-1503-28573) awarded to Michael J. Constantino. The statements in this article are solely the responsibility of the authors and do not necessarily represent the views of PCORI, its Board of Governors, or Methodology Committee. This study was also supported by a Division 29 (The Society for the Advancement of

Psychotherapy) of the American Psychological Association Grant awarded to James F. Boswell.

David R. Kraus is the founder, president, and chief scientific officer of Outcome Referrals, Inc., which owns and processes the Treatment Outcome Package. The authors have no other conflicts of interest to disclose.

The present article is original and has not been published elsewhere. There are two previous publications that analyzed data from overlapping patient and therapist samples.

The first publication (MS1) reported on the results of a randomized controlled trial that compared the effectiveness of prospectively matching patients to therapists' historical problem-based strengths to case assignment as usual. To

continued

As a maturing focus in psychotherapy practice and research, routinely administered patient-reported outcome measures can yield useful information about a given patient's treatment response and also a given therapist's general effectiveness across their cases. Regarding the latter, data reveal robust differences between clinicians in their average patient's treatment outcomes—the so-called between-therapist effect. This effect manifests reliably across different treatments administered in both controlled clinical trials (for which therapist actions are largely standardized) and routine care (for which therapist actions are largely unstandardized; Nissen-Lie et al., in press; Wampold & Owen, 2021). Moreover, between-therapist effects can exist both for global symptomatic/functional outcomes and more circumscribed outcomes (e.g., posttraumatic stress, depression), and they can hold even when adjusting for patient case-mix factors known to influence outcomes (Coyne, in press). Additionally, these therapist differences are clinically meaningful. According to systematic reviews (Johns et al., 2019) and meta-analyses (Baldwin & Imel, 2013), the therapist explains an average of about 5% of variance in patient outcomes, with this effect being notably higher for patients with certain types of presenting problems (e.g., problems in close relationships; Nissen-Lie et al., 2016) or with more versus less severe presenting concerns overall (Johns et al., 2019).

A smaller literature has also examined patient outcomes multidimensionally in order to determine whether therapists possess relative effectiveness strengths and weaknesses in treating their average patient with different types of presenting problems—so-called within-therapist effects. To date, the results of such research are somewhat mixed. For example, several studies have demonstrated that therapists who were effective (or ineffective) in one outcome domain also were similarly effective (or ineffective) in other domains (e.g., Green et al., 2014; Nissen-Lie et al., 2016). Notably, these studies focused on globally measured outcome

domains, such as social role performance, problems in general functioning, and physical symptoms. In contrast, when assessing more specified outcome domains (e.g., depression, substance misuse, suicidality), several other studies indicated that many therapists do show problem-specific effectiveness strengths and weaknesses (Constantino et al., 2021; Kraus et al., 2011, 2016), thereby reflecting a multidimensional performance profile or “report card” (see Coyne, in press). Importantly, these contrasting findings on within-therapist effects can arguably coexist. It is plausible that some therapists can simultaneously be consistently more or less effective than other clinicians when treating broadly defined outcome domains and yet still be differentially effective when treating more specified outcomes that may more closely resemble features of diagnosable clinical syndromes.

When such measurement-based, within-therapist strengths and weaknesses do exist, they also appear to be stable over time. That is, several studies have demonstrated that past therapist effectiveness in a given domain largely predicts future effectiveness in the same domain (Constantino et al., 2021; Kraus et al., 2016). Accordingly, taking a measurement-based approach to understanding therapist problem-specific performance profiles could help clinicians more accurately advertise their areas of historical effectiveness and, at the same time, assist patients in finding a therapist with the empirical track record that enhances their personal likelihood of being optimally helped. Currently, however, most publicly available therapist directories (e.g., Psychology Today, <https://GoodTherapy.org>) rely solely on therapists' non-measurement-based self-assessments of their problem-specific practice strengths (e.g., checking boxes for having expertise in treating presentations like depression, anxiety, sleep issues). Unfortunately, to the extent such self-assessments were inaccurate or unreliable, they would present a notable public health concern—a point to which we return momentarily.

“prime” the match condition, the trial involved a baseline phase during which the researchers collected routine multidimensional, patient-reported outcomes data from at least 15 historical patients (i.e., nontrial participants) for each of 50 therapists. These data (from a total of 1,363 patients) were used to classify therapists' effectiveness strengths and weaknesses to which subsequent *trial* patients randomized to the experimental condition were matched. These are the same therapist and patient samples analyzed in the present study. Importantly, MS1 reported on the main match effect on the outcomes of 218 new, consenting *trial* participants, whereas the present study focused solely on the outcomes of the 1,363 patients from the baseline (pretreatment) phase who were treated by the 48 therapists plus two others who ended up not participating in the trial. Therefore, the outcomes data in the present study have not been reported previously, and they have only been indirectly used to prime the match algorithm in the MS1 trial.

The second publication (MS2) was required by the trial's funder—the Patient-Centered Outcomes Research Institute (PCORI). This final research report went through a peer-review process and was published online by PCORI themselves. As part of this report, the researchers reported in an ancillary appendix on a preliminary analysis (consistent with the second aim of the present study) of a subset of the present study's therapists. With the fuller sample, the current authors reran these analyses, which are reported in the present article next to Aims 1 and 3, which were not presented in the funder-published final research report.

Michael J. Constantino played lead role in conceptualization, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, writing—original draft and writing—review and editing, and equal role in data curation and formal analysis. James F.

Boswell played supporting role in data curation, formal analysis, validation, and writing—original draft and equal role in conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, and writing—review and editing. Alice E. Coyne played lead role in formal analysis and visualization, supporting role in conceptualization, investigation, methodology, project administration, and writing—original draft and equal role in data curation, validation, and writing—review and editing. Heather J. Muir played supporting role in conceptualization, formal analysis, methodology, validation, and writing—review and editing. Averi N. Gaines played supporting role in conceptualization, data curation, methodology, validation, and writing—review and editing. David R. Kraus played supporting role in formal analysis, validation, and writing—original draft and equal role in conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, and writing—review and editing.

The study was not preregistered. The authors will maintain their de-identified data set indefinitely given possible future requests (e.g., to extract data for meta-analyses). Upon request on a case-by-case basis, they can share individual participant data (again, following full de-identification) that underlie results in a publication. Review requests, including a detailed plan for how the data will be used, should be directed via email to Michael J. Constantino (constanm@umass.edu). Michael J. Constantino reserves the right to deny individual share requests if the usage details are questionable or unclear.

Correspondence concerning this article should be addressed to Michael J. Constantino, Department of Psychological and Brain Sciences, University of Massachusetts Amherst, 135 Hicks Way, Amherst, MA 01003-9271, United States. Email: constanm@umass.edu

As noted, to establish therapists' measurement-based performance strengths and weaknesses in treating specified domains requires data from a multidimensional and routinely administered patient-reported outcomes measure, such as the Treatment Outcome Package (TOP; Kraus et al., 2005). As further described in the measures section, the TOP dimensionally assesses the following 12 specific and psychometrically determined outcomes: depression, quality of life (QOL), mania, panic/anxiety, psychosis, substance misuse, social conflict, sexual functioning, sleep, suicidality, violence, and work functioning. To arrive at therapist effectiveness classifications for each domain, researchers have used a multistep method (see Constantino et al., 2021; Kraus et al., 2011, 2016).

First, in a reference sample of approximately 28,000 adults receiving naturalistic psychotherapy in outpatient clinics, hospitals, residential centers, or day treatments, a machine learning analysis was used to determine baseline patient factors (e.g., demographic identities, symptom severity, and number of life stressors) that predicted TOP-based change from pre- to posttreatment (see Kraus et al., 2011, 2016). For each TOP domain, the resulting best-fitting model generated normative, expected change rates when accounting for the patient case-mix factors known to influence domain-specific treatment outcomes. Second, for any new patient who was not in the reference sample but was now receiving treatment in a similar community-based setting, the machine learning algorithm compares their personally expected rate of change, based on the case-mix-adjusted normative data established in the first step, to their actual rate of change in each TOP domain. Finally, for any therapist who has treated multiple cases with pre-posttreatment TOP data, a confidence interval (CI) can be calculated for their average patient's difference from the case-mix-adjusted, expected change rate for each clinical domain. Resulting from this analysis, therapists can be classified on each TOP domain as "effective" (meaning their average patient reliably exceeded their expected improvement), "neutral" (meaning their average patient changed to roughly the expected degree), or "ineffective" (meaning their average patient fell short of their expected improvement; Kraus et al., 2011, 2016). With these classifications established, a given therapist has a 12-domain effectiveness profile, or report card, that can be used to inform a fuller version of evidence-based treatment.

In one such application, researchers conducted a double-masked randomized clinical trial that compared the effectiveness of prospectively matching patients to therapists with a known track record of effectiveness in treating the patient's most salient presenting problem(s) to case assignment as usual (CAU), which is typically based on pragmatic considerations and leaves being matched empirically to chance (Constantino et al., 2021). Following this manipulated case assignment, treatment itself was delivered fully naturalistically in a large community mental health care system. As predicted, matched patients ($n = 99$) had greater symptom reduction and functional improvement than CAU patients ($n = 119$) across up to 16 weeks of treatment (with a medium-to-large effect size of $d = 0.75$). On its own, this system of matching patients to therapists' problem-specific effectiveness strengths has clear clinical promise; however, using therapist-level data in this manner would seem especially important if therapists were unable to make accurate self-assessments of their actual patient-reported, measurement-based effectiveness in treating different mental health concerns (Muir et al., 2019).

Speaking to this question, some research has suggested that inaccurate therapist self-assessments are likely the norm. For

example, one study showed that therapists tended to overrate their cognitive therapy competence relative to an expert independent rater (e.g., Brosan et al., 2008). In another study, therapists rated themselves and their teams as being better clinicians than their peers (Parker & Waller, 2015). They also reported exceptionally positive therapy outcomes. These results align with a high-profile survey for which 91.6% and 100% of therapist respondents, respectively, rated themselves as being in the top quartile and top half of effectiveness among their peers (Walfish et al., 2012). Moreover, 0% rated themselves as below average. These statistical impossibilities further demonstrate the seemingly powerful degree to which therapists view their abilities in what may be an overly positive light—a type of perceptual bias. Of course, this self-rated survey study did not include the measurement of actual therapist-level outcomes in their clinical practice. To assess more rigorously therapist accuracy in perceiving their clinical performance would require (a) examining the association between therapists' self-perceptions of their effectiveness and their actual measurement-based effectiveness and (b) doing so by specific presenting-problem domain in light of the aforementioned research demonstrating that many therapists possess some effectiveness strengths and some effectiveness weaknesses.

With such data, one can also assess whether therapists' self-perceptions of their strengths and weaknesses (i.e., how they view their own within-therapist effects) predict their actual overall effectiveness in helping patients improve on a global symptomatic/functional index (i.e., the measurement-based between-therapist effect). To this end, we drew on the TOP and community-based therapist and patient samples to address three unique aims. First, we assessed descriptively whether therapists possessed any perceptual biases; that is, we calculated the percentage of therapists in the sample who accurately, under-, or overestimated their own measurement-based effectiveness on each of the TOP domains. Second, we tested for the presence of therapist perceptual bias inferentially; that is, we examined how accurate therapists were at predicting to a level greater than chance their own objective effectiveness classification on each of the TOP domains (inaccuracy would suggest perceptual bias). Finally, we tested whether therapists' overall accuracy or inaccuracy of their effectiveness perceptions (across the 12 TOP domains) predicted actual between-therapist differences in their global effectiveness. As a sub-aim of this question, we also examined whether the association was stronger for individuals with more severe presenting problems (i.e., a cross-level interaction), as prior research has demonstrated that global therapist effects tend to be more pronounced for such patients (Johns et al., 2019). To our knowledge, this was the first study to focus on any of these aims; thus, the analyses were exploratory in nature.

Method

Data Set Overview

Data for this study derived from the baseline, prerandomization phase of the aforementioned controlled trial that tested the effectiveness of an empirical match system versus CAU prior to naturalistic outpatient therapy in a private community mental health system in the midwestern United States (Constantino et al., 2021). (To be sure, the trial provided methodological infrastructure; however, the distinct present study was fully confined to data collected at the trial's baseline.) As part of their standard clinical procedures, the six participating clinics had been routinely collecting patient pre- and

posttreatment TOP data prior to the trial. This historical archive allowed the researchers to classify the enrolled therapists' pretrial effectiveness report cards (as per the multistep procedure outlined previously) that ultimately allowed new trial patients in the experimental condition to be prospectively assigned to an empirically good-fitting clinician. For the present study's aims, these historical archived patient TOP data were analyzed at the therapist level, as described below. More specifically, they served as the measure of therapist effectiveness both in terms of problem-domain-specific strengths and weaknesses (i.e., the within-therapist effects at the center of Aims 1 and 2) and global effectiveness (the between-therapist effect at the center of Aim 3).

Participants

This study included 50 licensed therapists (14 psychologists, 36 masters-level clinical counselors/social workers) who were enrolled in the match trial and had an archive of baseline pre- and posttreatment TOP data on at least 15 patients¹ ($M = 27.26$; $SD = 4.08$; range = 16–30) to establish their historical problem-specific performance classifications for each of the 12 TOP domains. These clinicians averaged 49.46 years of age ($SD = 15.33$), and the majority identified as White (88%) and as a woman (74%). In terms of postlicense experience, they averaged 16.88 years ($SD = 12.39$). Across the 50 therapists, the archived patient sample included 1,363 adults. These patients averaged 36.71 years of age ($SD = 14.33$), and the majority identified as White (74%)² and as a woman (66%). As is typical for community outpatient mental health care, patients presented with diverse problems. The following data reflect the percentages of the sample who had clinical elevations (i.e., ≥ 1 SD above the general, non-treatment-seeking population) on each TOP domain, listed from most to least common: QOL (73%), depression (72%), panic/anxiety (48%), sleep (44%), social functioning (42%), psychosis (35%), suicidal ideation (30%), work functioning (28%), sexual functioning (28%), substance misuse (26%), mania (10%), and violence (9%).

Treatment

The archived patient-reported TOP data were collected in the context of naturalistic therapy of varied lengths. For the purposes of the match trial's baseline and the present study, the posttreatment assessment was the final completed follow-up TOP up to a maximum of 26 weeks (i.e., ~6 months). Within this outermost limit, the average length of treatment was 10.50 weeks ($SD = 5.56$). Although the exact nature of treatment was unknown, the therapists did report (at the trial's baseline) the degree to which their practice was influenced by major psychotherapy orientations. From highest to lowest—on a rating scale of 0 (*not at all*) to 6 (*very much*)—their mean ratings were cognitive behavioral ($M = 5.08$; $SD = 1.12$), integrative ($M = 4.36$; $SD = 1.55$), interpersonal ($M = 4.07$; $SD = 1.37$), humanistic/experiential ($M = 3.48$; $SD = 1.61$), systems ($M = 3.16$; $SD = 1.46$), and psychodynamic/psychoanalytic ($M = 2.57$; $SD = 1.76$).

Measures

Patient-Reported Outcomes

All indices in this study were derived from, or in relation to, the TOP (Kraus et al., 2005)—a routine measure of symptomatic/

functional impairment that historical patients completed as part of their standard care. Each of the TOP's 58 items is rated from 0 (*none*) to 5 (*all*) to capture how much time over the past 2 weeks the person has experienced a specific concern. As noted, the items load onto 12 clinical domains. For analysis and ease of clinical interpretation, these subscale data are transformed into z scores (i.e., SD units relative to the general, non-treatment-seeking population mean), with higher scores indicating greater impairment (e.g., a score of 3 on the depression domain would represent a depression level that is 3 SD s above the general population norm). The TOP has strong and well-established psychometric properties; the subscales have excellent factor structure, good internal consistency, and good test-retest reliability, and the total score (i.e., average of the 12 z scores) has excellent reliability, convergent validity, and change sensitivity (see Kraus et al., 2005).

Therapist Measurement-Based Effectiveness

For study Aims 1 and 2, which focused on perceptual bias, we assessed therapists' measurement-based effectiveness with the previously described multistep method for classifying therapists' problem-specific strengths and weaknesses (Constantino et al., 2021; Kraus et al., 2011, 2016). To reiterate, therapists were classified on each of the 12 TOP domains as effective (i.e., their average sample patient reliably exceeded their case-mix-adjusted expected outcome for the domain, as determined with machine learning in a large, separate reference sample), neutral (i.e., their average patient neither exceeded nor fell short of their case-mix-adjusted expected outcome for the domain), or ineffective (i.e., their average patient fell short of their case-mix-adjusted expected outcome for the domain). For study Aim 3, which focused on therapist self-perceived effectiveness as a predictor of their actual measurement-based between-therapist outcome differences, we used the TOP total score as the index of global outcome. That is, the therapists' caseload average total TOP score at posttreatment was the dependent variable, and their caseload average total TOP score at pretreatment was a covariate.

Therapist Self-Perceived Effectiveness

For all aims, we assessed therapists' self-perceived effectiveness with the Therapist Perceived Strengths (TPS) questionnaire. We designed the TPS to correspond with the patient-reported TOP; that is, therapists rated their perceived effectiveness on the same 12 outcome domains that are used to assess therapists' TOP-based effectiveness. Sample items included: "In treating my clients' symptoms of [panic/anxiety], I would say that I am ..."; "In improving my clients' [social functioning], I would say that I am:" The scale was as follows: 1 (*always ineffective*), 2 (*usually ineffective*), 3 (*sometimes ineffective*), 4 (*inconsistently ineffective*), 5 (*sometimes effective*), 6 (*usually effective*), and 7 (*always effective*). To determine the extent to which therapists' TPS ratings varied across outcome domains (which would reflect their ability to see both relative performance strengths and weaknesses, which was our construct of interest), we calculated an

¹ An average of 27 patients per therapist is consistent with the recommended sample size for reliably estimating therapist effects (Schiefele et al., 2017).

² Note that 179 patients (13%) chose not to provide race/ethnicity data.

intraclass correlation (ICC) with outcome domain self-perceptions nested within therapists. Results indicated that 91.6% of the variance was due to within-therapist variability, suggesting that therapists do perceive themselves as having performance strengths and weaknesses that the TPS can adequately capture.

In order to have full comparative alignment, TPS responses of 6 and 7 were collapsed to map onto the TOP-based performance classification of effective. TPS responses of 3, 4, and 5 were collapsed to map onto the TOP-based performance classification of neutral. Finally, TPS responses of 1 and 2 were collapsed to map onto the TOP-based performance classification of ineffective. With these categories, we could compare therapists' self-perceived effectiveness with their measurement-based strengths and weaknesses for Aims 1 and 2. For Aim 3, we derived three continuous accuracy variables to test as predictors of between-therapist differences in their global effectiveness as per the TOP total score. The first variable, *accuracy*, reflected the number of TOP domains (possible range of 0–12) for which therapists' perceived effectiveness classification matched their measurement-based effectiveness classification. The second variable, *underestimation*, reflected the number of TOP domains (0–12) for which therapists' perceived effectiveness classification was lower than their measurement-based effectiveness classification (i.e., therapists saw themselves as ineffective or neutral in treating a given problem when they were actually neutral or effective, respectively). Finally, the third variable, *overestimation*, reflected the number of TOP domains (0–12) for which therapists' perceived effectiveness classification was higher than their measurement-based effectiveness classification (i.e., therapists saw themselves as effective or neutral in treating a given problem when they were actually neutral or ineffective, respectively).

Procedure

Therapists within the community network were recruited for the match trial through emails or telephone calls (see Constantino et al., 2021, for additional details). At the trial's baseline, the consenting therapists completed a survey of measures that included the TPS, as well as questions focused on demographics, clinical training and experience, and influences from theoretical orientations. Also relevant to the present study, routinely collected, de-identified pre- and posttreatment patient TOP data for at least 15 of each participating therapist's historical cases were used to inform therapists' measurement-based, multidimensional effectiveness classifications. The institutional review board (protocol no. 2016-3401) at the trial's primary university approved the trial itself, as well as the additional analysis of de-identified data (including, for this study, exclusively at the therapist level).

Data Analyses

To address our first study aim, we calculated the percentage of therapists who over-, under-, or accurately estimated their measurement-based effectiveness classification for each of the 12 TOP domains.³ Addressing our second study aim, we inferentially examined therapist perceptual bias by conducting chi-square analyses to test whether therapists predicted their own measurement-based effectiveness classifications to a degree better than chance. More specifically, for each TOP domain, we compared therapists' perceived performance classifications based on their TPS responses

(i.e., ineffective, neutral, effective) to their TOP-based effectiveness classifications (i.e., ineffective, neutral, effective).

To test our third aim, we used hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) to account for the nested data structure (patients within therapists). More specifically, we fit three (one for each of our accuracy variables) two-level models with between-patient (within-therapist) variability at Level 1 and between-therapist differences at Level 2. As noted, patients' posttreatment TOP total score served as the outcome variable, which was predicted by each of the three therapist-level accuracy variables (at Level 2);⁴ that is, we tested whether between-therapist differences in accuracy, underestimation, and overestimation predicted between-therapist differences in their average patient's posttreatment global outcome.

Additionally, we controlled for within-therapist (Level 1) and between-therapist (Level 2) differences in patients' global baseline symptomatic/functional impairment (TOP total score) and treatment length. To generate the within-therapist covariates, we group-mean centered both variables. To generate the between-therapist covariates, we used each therapist's average of these variables across all patients in their caseload. To explore the cross-level severity by therapist accuracy interactions, we allowed the Level-1 association between within-therapist baseline patient impairment severity and posttreatment outcomes to vary across therapists (i.e., a random slope), which was then predicted by the relevant therapist-level self-perceived effectiveness accuracy variable at Level 2, while controlling for any differences in caseload-level severity. We also included a random intercept to allow patients' posttreatment outcome to vary across therapists, and we based the inclusion of a random slope for the within-therapist association between treatment length and posttreatment outcome on a chi-square model comparison test that determined whether a model with versus without the relevant random effect was a significantly better fit to the data. See the online Supplement Material, for the full multilevel equation for the best-fitting model.

Finally, given that therapists who were less effective than others in a measurement-based sense may have simply had fewer opportunities to underestimate their effectiveness, whereas those who were more effective than others in a measurement-based sense may have had fewer opportunities to overestimate their effectiveness, it is possible that our Aim 3 associations would be spurious. To address this, we conducted two distinct sensitivity analyses. First, we replicated our underestimation model while controlling for the number of domains on which a given therapist had a measurement-based classification of ineffective. Second, we replicated our overestimation model while controlling for the number of domains on which a given therapist had a measurement-based classification of effective. Although including these covariates may have resulted in models that were overly conservative, we considered it important to explore whether the therapist self-perceived effectiveness variables related to outcome when controlling for each therapist's relative chance to under- or overestimate their own measurement-based effectiveness.

³ The study was not preregistered. The data set and additional analytic details are available from the corresponding author upon request.

⁴ Given that therapists' scores on each of the three self-perceived performance accuracy classifications were inherently correlated (due to the fact they represented a count of the self-perception accuracy across the 12 domains), we examined each classification in a separate model.

Results

Prior to addressing our first aim, we examined therapists' overall patterns of measurement-based, domain-specific effectiveness. Consistent with previous research (Kraus et al., 2011, 2016), therapists had an average of approximately 1–2 relative strengths (M number of TOP domains effective = 1.52; SD = 1.95; range = 0–9) and approximately 0–1 relative weaknesses (M number of domains ineffective = 0.52; SD = 0.95; range = 0–4). Moreover, therapists were generally not universally effective or ineffective; 90% had three or fewer strengths and 96% had three or fewer weaknesses. Therefore, in this sample, the majority of therapists had a relatively equal chance to over-, under-, or accurately estimate their own measurement-based, domain-specific effectiveness.

Aim 1: Descriptive Within-Therapist Perceptual Bias

For the following 7 of 12 TOP domains, at least half of the therapists (M = 72%; range = 50%–82%) overestimated their own effectiveness: anxiety (82%), suicidality (82%), quality of life (80%), depression (76%), work functioning (70%), social functioning (62%), and violence (50%). Among the remaining five domains, the majority of therapists (M = 58%; range = 54%–62%) accurately estimated their own effectiveness: psychosis (62%), sleep (60%), mania (58%), substance misuse (54%), and sexual functioning (54%). Notably, there were no domains for which the majority of the therapists underestimated their own effectiveness (range = 0%–26%). See Table 1, for the full descriptive breakdown of therapist domain-specific perceptual bias.

Aim 2: Inferential Within-Therapist Perceptual Bias

Chi-square analyses revealed that across 11 of the 12 TOP domains, therapists were no better than chance at predicting their own measurement-based effectiveness (all ps > .05). The only exception was for the psychosis domain, $\chi^2(4) = 10.14$, $p = .038$,

Table 1

Descriptive and Predictive Within-Therapist Perceptual Bias by TOP Outcome Domain (N = 50)

TOP domain	Descriptive perceptual bias			Predictive perceptual bias ^a
	Over	Accurate	Under	
Anxiety	82%	18%	0%	$\chi^2(2) = 0.59$, $p = .745$
Suicidality	82%	18%	0%	$\chi^2(2) = 0.71$, $p = .702$
Quality of life	80%	20%	0%	$\chi^2(2) = 1.30$, $p = .522$
Depression	76%	20%	4%	$\chi^2(4) = 0.85$, $p = .932$
Work functioning	70%	26%	4%	$\chi^2(2) = 1.58$, $p = .453$
Social functioning	62%	34%	4%	$\chi^2(2) = 2.38$, $p = .304$
Violence	50%	42%	8%	$\chi^2(2) = 0.28$, $p = .869$
Psychosis	12%	62%	26%	$\chi^2(4) = 10.14$, $p = .038$
Sleep	34%	60%	6%	$\chi^2(2) = 0.71$, $p = .702$
Mania	26%	58%	16%	$\chi^2(4) = 0.96$, $p = .612$
Substance misuse	26%	54%	20%	$\chi^2(4) = 3.75$, $p = .441$
Sexual functioning	34%	54%	12%	$\chi^2(2) = 0.17$, $p = .920$

Note. TOP = Treatment Outcome Package.

^aNote that the degrees of freedom for the chi-square analyses differ depending on whether any therapists fell into certain categories for a given domain (e.g., no therapists saw themselves as ineffective at treating quality of life).

Cramer's $V = 0.318$; for this domain, 79% of the therapists with a measurement-based neutral effectiveness classification accurately perceived themselves as neutral (i.e., 30 out of 38 therapists), whereas only 17% of the measurement-based ineffective therapists correctly classified themselves as such (i.e., 1 out of 6 therapists) and 0% of the measurement-based effective therapists correctly classified themselves as such (i.e., 0 out of 6 therapists). See Table 1, for the results of all chi-square analyses by domain.

Aim 3: Therapist Self-Perceived Effectiveness as a Predictor of Global Between-Therapist Effectiveness

Results of an unconditional HLM (i.e., without predictors) revealed that although only 1.45% of the variance in patients' posttreatment outcome (i.e., TOP total score) was accounted for by the therapist, this effect was statistically significant, $\chi^2(49) = 68.12$, $p = .036$. Furthermore, when accounting for between-patient differences in baseline impairment severity, therapists accounted for 4.69% of the unexplained variance in patients' posttreatment outcome, which again was statistically significant, $\chi^2(49) = 114.07$, $p < .001$. This significant variability suggested there were meaningful between-therapist differences in their global effectiveness that could be explained by the addition of predictors. Moreover, a multilevel reliability estimate suggested that therapists' caseload-level posttreatment outcomes were measured with sufficient precision to proceed with the predictor analyses (reliability estimate = 0.56; Raudenbush & Bryk, 2002).

Next, we added the other relevant covariates—within-therapist (Level 1) and between-therapist (Level 2) differences in patients' baseline impairment severity and treatment length—and conducted model comparison tests to determine the best fit. Results revealed that a model in which between-patient severity and treatment length were allowed to vary across therapists (i.e., random slopes) and between-therapist differences in caseload-level baseline impairment were included as a Level-2 covariate was the best fit to the data (all model comparison $ps < .05$). In contrast, because between-therapist differences in caseload-level treatment length did not significantly predict any of the Level-2 outcomes (that is, posttreatment outcome [intercept], within-patient severity-outcome association [slope], within-patient treatment length-outcome association [slope]) and including it did not improve model fit, $\chi^2(3) = 0.17$, $p > .500$, we did not include it as a covariate for our primary analyses.

Our primary Aim 3 analyses showed that therapist accuracy was unrelated to between-therapist differences in their caseload-level posttreatment outcome ($\gamma_{02} = -0.01$, $SE = 0.01$, $p = .433$; 95% CI $[-0.03, 0.01]$; pseudo $r^2 = 0.037$), it did not moderate the within-therapist initial severity-outcome association ($\gamma_{12} = -0.03$, $SE = 0.02$, $p = .093$; 95% CI $[-0.06, 0.003]$), and adding it did not significantly improve model fit ($\chi^2[3] = 2.85$, $p > .500$). However, greater therapist underestimation was associated with better caseload-level posttreatment outcome ($\gamma_{02} = -0.06$, $SE = 0.02$, $p < .001$; 95% CI $[-0.10, -0.03]$; pseudo $r^2 = 0.48$). In terms of effect size, underestimation resulted in a 48% reduction in the unexplained variance in caseload-level posttreatment outcome. Putting this in terms of the overall therapist effect, which was 2% after accounting for the effects of all covariates, underestimation explained half (1%) of the unexplained variance in the therapist effect. Underestimation also significantly moderated the within-therapist severity-outcome association ($\gamma_{12} = -0.05$,

$SE = 0.02, p = .046$; 95% CI $[-0.09, -0.002]$), and adding it significantly improved model fit ($\chi^2[3] = 12.99, p = .005$). Specifically, for patients who presented with higher baseline impairment severity, therapist underestimation was more strongly associated with greater patient improvement, whereas such underestimation appeared to have little-to-no impact on outcome for patients with less severe baseline impairment (see Figure 1, Panel A).

Finally, greater therapist overestimation was associated with worse caseload-level posttreatment outcome ($\gamma_{02} = 0.03, SE = 0.01, p = .009$; 95% CI $[0.01, 0.05]$; pseudo $r^2 = 0.33$). Putting this in terms of the overall therapist effect, which was 2% after accounting for the effects of all covariates, overestimation explained 0.73% of the unexplained variance in the therapist effect. Overestimation also significantly moderated the within-therapist severity–outcome association ($\gamma_{12} = 0.04, SE = 0.01, p = .009$; 95% CI $[0.01, 0.06]$), and adding it significantly improved model fit ($\chi^2[3] = 10.32, p = .016$). Specifically, for patients who presented with higher baseline impairment severity, therapist overestimation was more strongly associated with worse posttreatment outcomes, whereas such overestimation had little-to-no impact on posttreatment outcome for patients with less severe baseline impairment (see Figure 1, Panel B). For the full results of the Aim 3 HLMs, see Supplemental Table 1.

Although our descriptive data indicated the majority of therapists had a relatively equal chance of over-, under-, or accurately estimating their own measurement-based, domain-specific strengths and weaknesses, we still conducted our previously described sensitivity analyses. As noted, we first replicated our underestimation model controlling for the total number of TOP domains for which therapists had a measurement-based classification of ineffective. Results indicated that greater underestimation was still associated with better caseload-level posttreatment outcome ($\gamma_{02} = -0.04, SE = 0.02, p = .011$; 95% CI $[-0.08, -0.01]$), and it still moderated

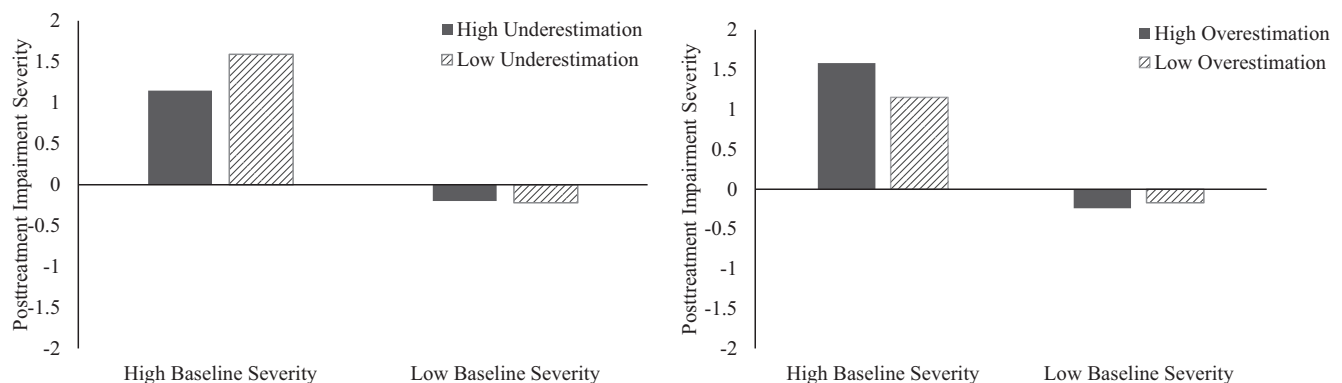
the within-therapist severity–outcome association; underestimation was associated with better posttreatment outcome for patients with more severe baseline impairment ($\gamma_{12} = -0.05, SE = 0.02, p = .046$; $-0.09, -0.01$). Second, when controlling for the overall number of domains for which therapists had a measurement-based classification of effective, overestimation was no longer associated with caseload-level posttreatment improvement ($\gamma_{02} = 0.01, SE = 0.01, p = .505$; 95% CI $[-0.01, 0.03]$), though greater overestimation still moderated the within-therapist severity–outcome association; overestimation was associated with worse posttreatment outcome for patients with more severe baseline impairment ($\gamma_{12} = 0.04, SE = 0.01, p = .010$; 95% CI $[0.02, 0.06]$).

Discussion

Most therapists promote their services based on their non-measurement-based self-perceptions of their problem-specific strengths. Thus, an important question is whether the public can rely on these attestations in making treatment decisions. This study examined, both descriptively and inferentially, the accuracy of therapists' self-perceptions of their own measurement-based, problem-specific effectiveness, and whether these self-perceptions predicted measurement-based differences in their global effectiveness. As per the descriptive statistics, therapists demonstrated perceptual bias; that is, on a majority of 12 different outcome domains, half or more of the therapists overestimated their own measurement-based effectiveness. Moreover, in no domains did the majority of therapists underestimate their own measurement-based effectiveness. As per the inferential statistics, for all but one outcome domain (i.e., psychosis), therapists were no better than chance at identifying their measurement-based performance classification. Finally, the average patient of therapists who were more likely to overestimate

Figure 1

Interactive Influence of Therapists' Accuracy or Inaccuracy of their Effectiveness Perceptions and Patient Baseline Severity on Patient Treatment Outcome



Note. Panel A depicts the cross-level interaction between therapist underestimation of their domain-specific effectiveness and within-therapist differences in patient presenting impairment severity. The dark solid bars depict the differential posttreatment outcomes for therapists with a high degree (+1.5 SDs) of underestimation when they treated patients with high (left side of Panel A) and low (right side of Panel A) presenting severity. The light dashed bars depict the differential posttreatment outcomes for therapists with a low degree (−1.5 SDs) of underestimation when they treated patients with high (left side of Panel A) and low (right side of Panel A) presenting severity. Panel B depicts the cross-level interaction between therapist overestimation of their domain-specific effectiveness and within-therapist differences in patient presenting severity. Specifically, the dark solid bars depict the differential posttreatment outcomes for therapists with a high degree (+1.5 SDs) of overestimation when they treated patients with high (left side of Panel A) and low (right side of Panel A) presenting severity. The light dashed bars depict the differential posttreatment outcomes for therapists with a low degree (−1.5 SDs) of overestimation when they treated patients with high (left side of Panel A) and low (right side of Panel A) presenting severity.

their actual domain-specific effectiveness reported worse global outcome than the average patient of therapists who were more likely to under- or accurately estimate their actual effectiveness (though a sensitivity analysis suggested that caution is warranted when interpreting this effect). Conversely, the average patient of therapists who were more likely to underestimate their actual domain-specific effectiveness reported better global outcome than the average patient of therapists who were more likely to over- or accurately estimate their actual effectiveness. Moreover, both the detrimental effect of greater therapist overestimation and the beneficial effect of greater therapist underestimation were more pronounced for patients with more severe presenting concerns.

The descriptive results further support Walfish et al.'s (2012) survey findings, which strikingly suggested that therapists possess an overconfidence bias. However, in the present study, this bias was supported by actual patient-reported outcomes data in relation to therapists' perceptions of their measurement-based, problem-specific effectiveness. The present data also revealed some nuance in therapists' effectiveness self-perceptions. Namely, the overconfidence bias was most notable for what are arguably the most widespread presenting problems in outpatient mental health care, including the three most commonly elevated problems in the present sample—QOL deficits, depression, and anxiety/panic. Among the remaining domains (i.e., sleep disturbance, psychosis, sexual functioning, substance misuse, mania), which reflect problems that are arguably less prevalent in general outpatient care (perhaps because they are viewed as requiring specialty and/or higher levels of care), the majority of therapists accurately estimated their own effectiveness, irrespective of their actual, TOP-based performance classification of effective, neutral, or ineffective. Thus, preliminarily at least, mental health care stakeholders may need to be most cautious when therapists advertise non-measurement-informed expertise in areas most likely to be included in their personal menu of services. Conversely, it is possible that stakeholders can place greater trust in therapists' non-measurement-informed claims of expertise when they relate to rarer conditions seen in general outpatient settings, and/or ones often viewed as more intractable. For example, if a therapist lists mania as a practice focus, they may have a relatively accurate sense of their ability to help people with this problem.

Importantly, though, accuracy does not necessarily mean effectiveness, as a clinician could conceivably be accurate in seeing themselves as largely ineffective in treating mania. Although one might hope this clinician would remove this focus when advertising their practice, it is plausible they retain it because they view this as the norm for all clinicians who treat this challenging condition. This scenario underscores the importance of using benchmarked, patient-reported outcomes tools to quantify therapist effectiveness and help remove some of the high-stakes guesswork (Rousmaniere et al., 2020). With such quality data comes the hope that stakeholders making mental health care decisions can make and receive more trusted referrals and case assignments to therapists' actual measurement-based strengths, including (to stick with the present example) to a clinician who may historically excel at treating mania versus one who just believes that most therapists do not perform particularly well in this area and therefore they are "better than nothing" and "similar to the next" (when, in fact, measurement-based outcomes could reveal them as *harmful* in this area; Boswell et al., 2022).

In addition to these sample-based descriptive results highlighting a partial overconfidence trend, it is important to spotlight the general predictive inaccuracy of therapist's performance self-perceptions. That is, only for the domain of psychosis were therapists able to predict their own measurement-based effectiveness to a significant degree. As noted previously, using therapist-level data to inform referrals and case assignments to empirically good-fitting clinicians takes on special importance if therapists have difficulty making accurate self-assessments of their problem-specific effectiveness, as the present data suggest they do. With more widespread replication of such inaccuracies, mental health care stakeholders would be right to question the veracity of therapists' self-marketing of effectiveness in the absence of data. Instead, these stakeholders and the systems within which they work or seek treatment could focus on explicit and data-driven personalization efforts that center on the provider versus the more typical focus on the patient or treatment type (Constantino & Muir, in press). As previously reviewed, a randomized trial (Constantino et al., 2021) indicated that prospectively matching patients to therapists with historical, data-informed strengths in treating the patient's most salient problem(s) significantly improved the outcome of subsequent naturalistic therapy when compared to usual case-assignment methods (which, again, were generally based on pragmatic considerations, including therapists' potentially inaccurate self-defined areas of effectiveness).

Importantly, even as poor prognosticators of their own problem-specific effectiveness, the direction of therapist self-assessments may have clinical meaning. Namely, it appears the most globally effective therapists, in relation to their peers, are the ones who do not consistently overestimate their own measurement-based effectiveness and instead somewhat underestimate it. In fact, our sensitivity analyses revealed therapist underestimation to be the most robust predictor of global between-therapist effectiveness differences, even when controlling for the number of domains for which therapists had a measurement-based classification of ineffective. This result adds to the currently limited literature on therapist-level factors that predict the between-therapist effect (Wampold & Owen, 2021). Moreover, it aligns with and extends one of the few variables to date that has emerged as a predictor of globally better versus worse performing clinicians—having more professional self-doubt in the form of critically questioning one's clinical skills (Nissen-Lie et al., 2013, 2017).⁵ Although these findings may appear counterintuitive, the beneficial impact of professional self-doubt (whether in the form of self-criticalness or underestimating one's measurement-based effectiveness strengths and weaknesses) may make sense if, instead of simply reflecting a negative self-perception, it is potentially tapping into the broader virtue of professional humility.

Reflecting the adaptive ability to maintain a balanced and accurate (or even somewhat cautious/modest) view of one's strengths and weaknesses, in this case in a therapist's inherently challenging clinical practice, humility has long been discussed as an important virtue for helping others (e.g., Kierkegaard, 1998). For example, starting with the broader positive psychology literature, possessing more humility has been associated with engaging in acts of generosity (Exline & Hill, 2012), initiating helpful behaviors toward

⁵ Whereas the Nissen-Lie et al. (2013, 2017) studies showed the effect of professional self-doubt on patient interpersonal outcomes only, the present results extend this effect to a more global symptomatic/functional outcome index.

strangers (LaBouff et al., 2012), and fostering a romantic partner's relationship satisfaction and commitment (Dwiwardani et al., 2018; Farrell et al., 2015). More specific to helping professions, greater physician humility has been positively associated with patients' beneficial health outcomes (Huynh & Dicke-Bohmann, 2020; Ruberton et al., 2016), reports of effective physician–patient communication (Ruberton et al., 2016), and both satisfaction with and trust in their physician (Huynh & Dicke-Bohmann, 2020). In a dyadic psychotherapy study, therapists who tended to rate therapeutic change less positively than their patients rated their own change were generally more versus less effective (Ziem & Hoyer, 2020). Taken together, a personally humbler attitude of one individual can have a variety of beneficial effects on others, including in the psychotherapy context. As just one speculation on its clinical mechanism, perhaps a degree of doubt in one's clinical abilities enables therapists to be more alert to challenges or signals that a given course of therapy (and the patient–therapist working relationship that underlies it) is off-track, thereby allowing for timely and flexibly responsive interventions that have a greater likelihood of success (Constantino et al., 2020; Macdonald & Mellor-Clark, 2015).

Furthermore, our moderator analyses revealed that the beneficial influence of therapist underestimation of their measurement-based strengths and, especially, the potentially damaging influence of therapist overestimation of their measurement-based strengths on therapist-level treatment outcome were magnified for patients with more severe baseline impairment and less pronounced for patients with less severe impairment. On the one hand, given that patients with more severe problems may be at greater risk for a difficult treatment course, it seems plausible that therapists who espouse a humbler and more vigilant stance could be more alert and responsive to such challenges. Alternatively, perhaps patients with more severe problems are particularly sensitive to therapist overconfidence and find it especially off-putting. For example, patients who have dealt with their mental health problems for a long time and have likely been in therapy before may find it invalidating for therapists to be unrealistically confident in their ability to help. On the other hand, the fact that therapist under- and overestimation were less influential for patients with less severe presenting concerns aligns with previous findings that between-therapist performance differences are less pronounced for these patients (e.g., Johns et al., 2019). Thus, the present results further suggest that individuals with less severe problems may be likely to improve when treated by most therapists, even if those therapists are less effective than their peers in a measurement-based sense and/or are poor judges of their measurement-based, problem-specific effectiveness.

These findings related to therapist under- and overestimation of their own effectiveness may also have some preliminary training implications. With further replication as determinants of global between-therapist effectiveness differences, it would become important to examine if therapist humility is mutable and can therefore be cultivated in clinical trainings. If so, it may be especially useful to use deliberate practice methods—one of the other factors that has emerged to date as a predictor of generally more versus less effective clinicians (e.g., Chow et al., 2015). When combined with using patient-reported outcomes data to inform therapists of their problem-specific strengths and weaknesses, we can envision trainings that take on both universal and personalized forms. Namely, it may be that all therapists could benefit from trainings that cultivate professional humility, whereas any given therapist can draw on their

performance report cards to personally train to reinforce their strengths and/or redress their weaknesses (see Boswell et al., 2022; Coyne, *in press*).

Of course, any implications of the present work should be considered with the study's limitations in mind. First, because the TPS was a new measure that we developed specially for this line of research, it potentially has reliability issues. For example, it is currently unknown whether therapists would demonstrate test–retest reliability in how they rate their own domain-specific effectiveness. Additionally, the frequency with which a therapist feels they encounter a certain problem dimension could affect how they respond to the TPS items. However, the TPS has high face and ecological validity for this study; that is, as previously noted, it mimics how therapists currently tend to advertise their domain-specific strengths without the influence of measurement. This standard-practice ecological validity (through platforms such as *Psychology Today* and <https://GoodTherapy.org>) was a significant motivation for the development of this particular tool.

Second, our assessment of therapists' self-perceived accuracy of their measurement-based, problem-specific effectiveness classifications led us to tether our results to the humility construct. Of course, ours was only an indirect assessment of humility, and future research will need to assess it more directly with well-established self- and other-report instruments. Third, regarding our Aim 3, it remains possible that more globally effective therapists simply had more opportunities to underestimate their own effectiveness and fewer opportunities to overestimate it, thereby rendering the associations between these variables and patient outcomes somewhat of a methodological artifact. However, this concern was partially mitigated by descriptive and sensitivity analyses. Fourth, we have no information about the factors on which therapists based their self-perceptions of effectiveness. Nevertheless, it is worth reiterating that the therapist sample was relatively seasoned, with an average of 16.88 years of postlicense experience to draw from when providing their TPS ratings.

Fifth, although this study focused on a multidimensional outcome measure that has historically revealed within-therapist differences in domain-specific effectiveness, we are aware that other research (that used different multidimensional outcome measures) has failed to show such differences (instead revealing that therapist effectiveness may be more of a global factor; e.g., Green et al., 2014; Nissen-Lie et al., 2016). Thus, the literature on within-therapist effects remains far from settled, and future research should continue to assess for which types of outcome dimensions therapists do or do not evidence within-person differences, as well as their ability to perceive their own general and/or differentiated effectiveness accurately. Finally, the study had restricted generalizability beyond a specific and relatively small sample of therapists in a mental health care system in the Midwestern United States. Generalizability was further limited by a relative lack of diversity in both the therapist and patient samples, as well as the fact that clinicians self-selected to participate in the research.

Limitations aside, as the field continues to learn more about therapist effects, and therapist-level factors that predict them (like humility), there is clear promise to make such findings clinically actionable (Constantino et al., 2021; Constantino & Muir, *in press*). Ultimately, we see the provider as an important, long-neglected, and nuanced aspect of what it means to be engaging in evidence-based mental health care.

References

- 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). Wiley.
- Boswell, J. F., Constantino, M. J., & Coyne, A. E. (2022). What works in therapy when delivered by whom? *Clinical Psychology: Science and Practice*, 29(2), 137–139. <https://doi.org/10.1037/cps0000072>
- Brosnan, L., Reynolds, S., & Moore, R. G. (2008). Self evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy*, 36(5), 581–587. <https://doi.org/10.1017/S1352465808004438>
- 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: Theory, Research, and Practice*, 52(3), 337–345. <https://doi.org/10.1037/pst0000015>
- Constantino, M. J., Boswell, J. F., Coyne, A. E., Swales, T. P., & Kraus, D. R. (2021). Enhancing mental health care by matching patients to providers' empirically derived strengths: A randomized clinical trial. *JAMA Psychiatry*, 78(9), 960–969. <https://doi.org/10.1001/jamapsychiatry.2021.1221>
- Constantino, M. J., Coyne, A. E., & Muir, H. J. (2020). Evidence-based therapist responsiveness to disruptive clinical process. *Cognitive and Behavioral Practice*, 27(4), 405–416. <https://doi.org/10.1016/j.cbpra.2020.01.003>
- Constantino, M. J., & Muir, H. J. (in press). Can we prospectively harness therapist effects for therapeutic benefit? In F. Leong, J. L. Callahan, M. J. Constantino, C. F. Eubanks, & J. Zimmerman (Eds.), *Handbook of psychotherapy*. American Psychological Association.
- Coyne, A. (in press). Therapist performance report cards: Do clinicians differ in their specific effectiveness? In F. Leong, J. L. Callahan, M. J. Constantino, C. F. Eubanks, & J. Zimmerman (Eds.), *Handbook of psychotherapy*. American Psychological Association.
- Dwiwardani, C., Ord, A. S., Fennell, M., Eaves, D., Ripley, J. S., Perkins, A., Sells, J., Worthington, E. L., Jr., Davis, D. E., Hook, J. N., Garthe, R. C., Reid, C. A., & Van Tongeren, D. R. (2018). Spelling HUMBLE with U and ME: The role of perceived humility in intimate partner relationships. *The Journal of Positive Psychology*, 13(5), 449–459. <https://doi.org/10.1080/17439760.2017.1291849>
- Exline, J. J., & Hill, P. C. (2012). Humility: A consistent and robust predictor of generosity. *The Journal of Positive Psychology*, 7(3), 208–218. <https://doi.org/10.1080/17439760.2012.671348>
- Farrell, J. E., Hook, J. N., Ramos, M., Davis, D. E., Van Tongeren, D. R., & Ruiz, J. M. (2015). Humility and relationship outcomes in couples: The mediating role of commitment. *Couple & Family Psychology*, 4(1), 14–26. <https://doi.org/10.1037/cfp0000033>
- 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. <https://doi.org/10.1016/j.brat.2014.08.009>
- Huynh, H. P., & Dicke-Bohmann, A. (2020). Humble doctors, healthy patients? Exploring the relationships between clinician humility and patient satisfaction, trust, and health status. *Patient Education and Counseling*, 103(1), 173–179. <https://doi.org/10.1016/j.pec.2019.07.022>
- Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review*, 67, 78–93. <https://doi.org/10.1016/j.cpr.2018.08.004>
- Kierkegaard, S. (1998). The point of view for my work as an author (E. H. Hong, & H. V. Hong, Trans.). In S. Kierkegaard (Ed.), *The point of view* (pp. 21–126). Princeton University Press. (Original work published 1859).
- Kraus, D. R., Bentley, J. H., Alexander, P. C., Boswell, J. F., Constantino, M. J., Baxter, E. E., & Castonguay, L. G. (2016). Predicting therapist effectiveness from their own practice-based evidence. *Journal of Consulting and Clinical Psychology*, 84(6), 473–483. <https://doi.org/10.1037/ccp0000083>
- Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research*, 21(3), 267–276. <https://doi.org/10.1080/10503307.2011.563249>
- Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology*, 61(3), 285–314. <https://doi.org/10.1002/jclp.20084>
- LaBouff, J., Rowatt, W., Shen, M., Tsang, J.-A., & Willerton, G. (2012). Humble persons are more helpful than less humble persons: Evidence from three studies. *The Journal of Positive Psychology*, 7(1), 16–29. <https://doi.org/10.1080/17439760.2011.626787>
- Macdonald, J., & Mellor-Clark, J. (2015). Correcting psychotherapists' blindsidedness: Formal feedback as a means of overcoming the natural limitations of therapists. *Clinical Psychology & Psychotherapy*, 22(3), 249–257. <https://doi.org/10.1002/cpp.1887>
- Muir, H. J., Coyne, A. E., Morrison, N. R., Boswell, J. F., & Constantino, M. J. (2019). Ethical implications of routine outcomes monitoring for patients, psychotherapists, and mental health care systems. *Psychotherapy: Theory, Research, and Practice*, 56(4), 459–469. <https://doi.org/10.1037/pst0000246>
- Nissen-Lie, H. A., Goldberg, S. B., Hoyt, W. T., Falkenström, F., Holmqvist, R., Nielsen, S. L., & Wampold, B. E. (2016). Are therapists uniformly effective across patient outcome domains? A study on therapist effectiveness in two different treatment contexts. *Journal of Counseling Psychology*, 63(4), 367–378. <https://doi.org/10.1037/cou0000151>
- Nissen-Lie, H. A., Monsen, J. T., Ulleberg, P., & Rønnestad, M. H. (2013). Psychotherapists' self-reports of their interpersonal functioning and difficulties in practice as predictors of patient outcome. *Psychotherapy Research*, 23(1), 86–104. <https://doi.org/10.1080/10503307.2012.735775>
- Nissen-Lie, H. A., Oddli, H. W., & Heinonen, E. (in press). Do therapists differ in their general effectiveness? Therapist effects and their determinants. In F. Leong, J. L. Callahan, M. J. Constantino, C. F. Eubanks, & J. Zimmerman (Eds.), *Handbook of psychotherapy*. American Psychological Association.
- Nissen-Lie, H. A., Rønnestad, M. H., Høglend, P. A., Havik, O. E., Solbakken, O. A., Stiles, T. C., & Monsen, J. T. (2017). Love yourself as a person, doubt yourself as a therapist? *Clinical Psychology & Psychotherapy*, 24(1), 48–60. <https://doi.org/10.1002/cpp.1977>
- Parker, Z. J., & Waller, G. (2015). Factors related to psychotherapists' self-assessment when treating anxiety and other disorders. *Behaviour Research and Therapy*, 66, 1–7. <https://doi.org/10.1016/j.brat.2014.12.010>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE Publications.
- Rousmaniere, T., Wright, C. V., Boswell, J., Constantino, M. J., Castonguay, L., McLeod, J., Pedulla, D., & Nordal, K. (2020). Keeping psychologists in the driver's seat: Four perspectives on quality improvement and clinical data registries. *Psychotherapy: Theory, Research, and Practice*, 57(4), 562–573. <https://doi.org/10.1037/pst0000227>
- Ruberton, P. M., Huynh, H. P., Miller, T. A., Kruse, E., Chancellor, J., & Lyubomirsky, S. (2016). The relationship between physician humility, physician-patient communication, and patient health. *Patient Education and Counseling*, 99(7), 1138–1145. <https://doi.org/10.1016/j.pec.2016.01.012>
- Schiefele, A. K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., Kopta, M., Schulte, D., Saxon, D., Nielsen, S. L., & Lambert, M. J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Administration and Policy in Mental Health*, 44(5), 598–613. <https://doi.org/10.1007/s10488-016-0736-3>
- Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports*, 110(2), 639–644. <https://doi.org/10.2466/02.07.17.PR0.110.2.639-644>

- Wampold, B. E., & Owen, J. (2021). Therapist effects: History, methods, magnitude, and characteristics of effective therapists. In L. G. Castonguay, M. Barkham, & W. Lutz (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 301–330). Wiley.
- Ziem, M., & Hoyer, J. (2020). Modest, yet progressive: Effective therapists tend to rate therapeutic change less positively than their patients.

Psychotherapy Research, 30(4), 433–446. <https://doi.org/10.1080/10503307.2019.1631502>

Received August 31, 2022

Revision received January 5, 2023

Accepted February 19, 2023 ■