Journal of Counseling Psychology
2007, Vol. 54, No. 1, 32–39

0022-0167/07/$12.00 DOI: 10.1037/0022-0167.54.1.32

Therapist Effects in Outpatient Psychotherapy: A Three-Level Growth Curve Approach

Wolfgang Lutz
University of Berne and University of Trier

Scott C. Leon
Loyola University Chicago

Zoran Martinovich and John S. Lyons
Northwestern University Medical School

William B. Stiles
Miami University

Evidence suggests that a moderate amount of variance in patient outcomes is attributable to therapist differences. However, explained variance estimates vary widely, perhaps because some therapists achieve greater success with certain kinds of patients. This study assessed the amount of variance in across-session change in symptom intensity scores explained by therapist differences in a large naturalistic data set (1,198 patients and 60 therapists, who each treated 10–77 of the patients). Results indicated that approximately 8% of the total variance and approximately 17% of the variance in rates of patient improvement could be attributed to the therapists. Cross-validation and extreme group analyses validated the existence of these therapist effects.

Keywords: therapist effects, three-level growth curve model, outcomes management, expected treatment response

Supplemental materials: http://dx.doi.org/10.1037/0022-0167.54.1.32.supp

As Kim, Wampold, and Bolt (2006) cogently pointed out, psychotherapy research's questions and methods historically were derived from research in disciplines such as farming, education, and medicine, which emphasized the intervention over the interventionist. As a result, psychotherapy research has devoted substantial resources to developing and testing therapies and comparatively less to the role of therapists (Garfield, 1997). Despite this relative neglect, the provocative question "Are some therapists more effective than others?" has been studied with increasing frequency since Ricks's (1974) famous "Supershrink" study (e.g., Project MATCH Research Group, 1998). Reviews have sought to summarize the percentage of outcome variance accounted for by therapists (Crits-Christoph et al., 1991; Crits-Christoph & Gallop, 2006; Crits-Christoph & Mintz, 1991; Crits-Christoph, Tu, & Gallop, 2003; Elkin, 1999; Lambert & Okiishi, 1997; Luborksy et al., 1986; Shapiro, Firth-Cozens, & Stiles, 1989). For example, in a meta-analysis of 27 studies, Crits-Christoph and Mintz (1991) found that therapist effects ranged from 0% to 50%, with a mean of 8.6%. Huppert, Bufka, Barlow, Gorman, and Shear (2001), analyzing the data from the Multicenter Collaborative Study for the Treatment of Panic Disorders, reported therapist effects ranging from 1% to 18% depending on the outcome measure.

In a recent special section of the journal *Psychotherapy Research*, Elkin, Falconnier, Martinovich, and Mahoney (2006) and Kim et al. (2006) separately used multilevel modeling strategies to assess therapist effects using data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP; Elkin et al., 1989). The NIMH TDCRP was a randomized clinical trial comparing two prominent treatments for major depression; therapists were all experienced clinicians trained to conduct the manualized treatments (interpersonal therapy vs. cognitive–behavioral therapy) and were monitored to ensure fidelity.

Even though they used the same data set, Elkin et al. (2006) found no significant therapist effects, whereas Kim et al. (2006) found that 5%–10% of the variance in outcomes could be attributed to therapists. Commentaries by Soldz (2006) and Crits-Christoph and Gallop (2006) as well as rejoinders by Wampolt and Bolt (2006) and Elkin (2006) offered explanations for the discrepant findings and suggestions for how to improve the study of therapist effects in future research. It was pointed out that the NIMH TDCRP data set was not ideal to study therapist effects because it included only 17 therapists and only 4–11 patients per therapist. Elkin et al. (2006) noted similarly limited statistical power in previous studies of therapist effects. Our study addressed this limitation by using a large, naturalistic data set.

In an earlier study of therapist effects in a naturalistic database, Okiishi, Lambert, Nielsen, and Ogles (2003) examined variation in patient outcomes for 91 therapists over 1,841 patients. Using

Wolfgang Lutz, Department of Psychology, University of Berne, Berne, Switzerland, and Department of Psychology, University of Trier, Trier, Germany; Scott C. Leon, Department of Psychology, Loyola University Chicago; Zoran Martinovich and John S. Lyons, Department of Psychiatry and Behavioral Sciences, Northwestern University Medical School; William B. Stiles, Department of Psychology, Miami University.

This work was partially supported by Swiss National Science Foundation Grants PP001-102651 and 1114-064884.0 to Wolfgang Lutz. We are grateful for the statistical work of Bruce Briscoe.

Correspondence concerning this article should be addressed to Wolfgang Lutz, Department of Psychology, University of Berne, Muesmattstrasse 45, CH-3012 Berne 9, Switzerland. E-mail: wolfgang.lutz@psy.unibe.ch

hierarchical linear modeling (Raudenbush & Bryk, 2002, Okiishi et al. (2003) found substantial variation in outcomes across therapists after controlling for intake and demographic characteristics of patients, although their descriptive method did not specify the percentage of outcome variance explained by therapists. The therapists whose patients showed the most improvement had an average change rate 10 times greater than the sample's mean rate. The researchers could replicate their findings on a larger data set of over 5,000 patients (Okiishi et al., 2006). In another hierarchical linear modeling study of therapist effects in naturalistic settings, Wampold and Brown (2005) reported the results of a two-level model analysis using a large administrative database of patients who received therapy from a major national managed care organization and who completed psychometric assessments pre- and posttherapy. The therapist effects ranged from 5% to 8%, very similar to Kim et al.'s (2006) analyses of the TDCRP data set.

For statistically analyzing therapist effects, many researchers agree that therapists should be treated as a random variable in a multilevel modeling approach to adequately generalize findings (for a discussion of fixed and random models of therapist effects, see, e.g., Crits-Christoph & Mintz, 1991; Serlin, Wampold, & Levin, 2003; Siemer & Joormann, 2003). As Kim et al. (2006) pointed out, "If therapists are treated as fixed, the results are conditioned on the particular therapists included in the clinical trial, thus restricting the conclusions to only those particular therapists in the trial" (p. 162). If therapists are treated as a random variable, conversely, results can be generalized to a theoretical population of therapists similar to the therapists under investigation (Serlin et al., 2003).

Reflecting on their failure to find significant therapist effects, Elkin et al. (2006) suggested that the therapist effects reported in previous studies might have reflected the researchers' failure to treat therapists as a random factor or to include all three levels of nesting—sessions nested within patients, patients nested within therapists, and therapists. Most previous studies have not assessed patient progress at each session and have not included the session level (sometimes called *time*) in their models. Other criticisms of previous studies include the use of therapist-rated measures (confounding rater biases with therapist effects) or a failure to conduct inferential tests of therapist effects in the first place. In our study, we used a patient self-report outcome measure, gathered at each session, and a three-level growth curve approach. Our three-level model treated therapists as a random variable and also accounted for dependencies in the data due to nesting.

A further methodological issue involves how to treat therapist outliers (Elkin et al., 2006). For instance, the significant therapist effects in Project MATCH Research Group (1998) were mainly due to 1 outlier out of 54 therapists. Clearly, the importance of therapist effects is more limited if they are attributable to a very small number of extremely good or poor therapists. In our study, we modeled therapist effects for the entire sample of therapists and then again with outliers removed.

In summary, the aim of this study was to assess the amount of outcome variance attributable to therapists practicing therapy in real-world settings using a large repeated-measurement data set including 1,198 patients treated by 60 therapists. Our multilevel data-analytic approach incorporated patient intake variables, paralleling the Okiishi et al. (2003) study, in effect controlling for differences in therapists' caseloads on these variables. We applied a three-level growth curve approach to assess the amount of variance explained at all three levels of the nested data: sessions, patients, and therapists. To demonstrate the stability of results, we used a cross-validation procedure. Finally, following Elkin et al.'s (2006) concern that therapist effects may be attributable to outliers, we reanalyzed our data after excluding extremely good and poor therapists.

Method

Participants

The patient sample included 1,198 psychotherapy outpatients who began therapy above the typical range in self-reported symptom intensity ratings (described later). The patients' mean age was 36.4 ($SD = 9.5$); 73% were women; 59% were married, 24% were single, and 18% were separated, divorced, or widowed; 88% were White, 6% were African American, 0.6% were Asian, 4.5% were Hispanic, 0.6% were Native American, and 0.6% were of other ethnic identification. Seventy percent were employed full time, and 75% had some college education. These statistics are reasonably representative of psychotherapy outpatients in the United States (cf. Merrill, Tolbert, & Wade, 2003; Stirman, DeRubeis, Crits-Christoph, & Brody, 2003; Vessey & Howard, 1993). Diagnoses based on the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994) were available for 1,102 patients (92%): 25.9% ($n = 285$) had a primary diagnosis of major depressive disorder, 18.1% ($n = 199$) were diagnosed with dysthymic disorder, 16.6% had an adjustment disorder with mixed anxiety and depressed mood (an additional 7.6%, $n = 183$, had an adjustment disorder with depressed mood, and an additional 2.9%, $n = 32$, had an adjustment disorder with anxiety), 5.4% ($n = 59$) had a generalized anxiety disorder, and 3.4% ($n = 37$) had a panic disorder (an additional 2.5%, $n = 27$, had another anxiety disorder).

The therapist sample included 60 therapists in the national provider network of an American managed care company. All therapists had formal training and at least 1 year postqualification experience; 65% of the therapists were female. The therapists varied in professional background and theoretical orientation, and many were familiar with published treatment manuals; however, they were not required to follow a formal manualized protocol in the treatments we studied, and their profession and approach were not systematically recorded. Treatment duration was variable and not subject to strict time limits, although some therapists set time limits as part of their treatment strategy.

Each therapist serviced a substantial number of insured cases (ranging from 10 to 77 patients per therapist). An appendix prepared as an online supplement to this article gives further information about the caseload of each therapist, including the average intake score on our mental health assessment measure (described later), the average treatment duration (in sessions), and frequencies of patients within the primary Axis I diagnosis-related group, frequencies of patients meeting criteria for any personality disorder, and the average change in Mental Health Index (MHI) from first to last session in standard deviation units. These statistics can be found on the Web. Most therapists saw patients with affective disorders, but several had a focus on patients with adjustment, anxiety, or personality disorders. Twenty-five of the therapists had fewer than 15 patients in their caseload, and 15 therapists had at least 20 patients.

Measures

The Compass tracking system, originally called the Integra Outpatient Treatment Assessment system (Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lueger et al., 2001; Lyons, Howard, O'Mahoney, & Lish, 1997), is one of a number of comprehensive assessment batteries that has been used to measure progress in outpatient mental health treatment and includes both patient and clinician assessments of a range of relevant

outcomes. The tracking system includes a global outcome criterion, the MHI, which consists of the sum of the Subjective Wellbeing scale, the Current Symptoms scale, and the Current Life Functioning scale. Subjective Wellbeing is a 4-item scale on which patients rate their overall distress level, health, energy level, emotional adjustment, and life satisfaction. Current Symptoms is a 40-item scale on which patients rate the frequency of symptoms experienced over the past 2 weeks. It reflects the *DSM* (3rd ed., rev.; American Psychiatric Association, 1987) diagnoses of adjustment disorder, anxiety, bipolar disorder, depression, obsessive–compulsive disorder, phobia, and substance abuse. Current Life Functioning is a 24-item scale on which patients rate the extent to which emotional and psychological difficulties are interfering with functioning in six main areas (e.g., work, family, self-management). The MHI has an internal consistency of .87 and a (3–4 week) test–retest stability of .82. The average MHI has been shown to be significantly lower for psychotherapy patients than for nonpatients (Lueger et al., 2001).

For the present analyses, the MHI was converted so that higher scores indicated better mental health and transformed to T scores (mean of 50 and standard deviation of 10) on the basis of first-session norms from over 16,000 patients. MHI T scores below 60 are considered as more representative of a patient population than of a nonpatient population (i.e., outside the normal range; see Jacobson & Truax, 1991). In previous work, this cutoff has been used to identify patients for whom treatments were medically necessary. It has also been used to determine successful treatment (Howard et al., 1996).

Included in the tracking system were three patient ratings on anchored rating scales: (a) prior psychotherapy ("How much counseling or psychotherapy have you had in the past?"; rated from *none* to *more than one year*), (b) chronicity ("How long has the problem for which you are presently seeking treatment been a concern to you?"; rated from *less than one month* to *more than two years*), and (c) treatment expectations ("When you finish counseling or psychotherapy, how well do you feel that you will be getting along emotionally and psychologically?"; rated from *quite poorly—I will be barely able to manage to deal with things* to *very well—much the way I would like to*).

Also included was the therapist-rated Global Assessment Scale (GAS) (Endicott, Spitzer, Fleiss, & Cohen, 1976), which is a 100-point anchored rating scale (subsequently modified and included in the *DSM–IV* as Axis V). The GAS and MHI correlated .45; the correlations among the other predictors were all below .20.

Procedure

The tracking system was applied to patients from a diverse national sample of therapists, settings, and psychotherapy patients, mostly from the eastern part of the United States, whose treatment was being managed with the assistance of the Compass System. The psychometric data were gathered primarily for and used by therapists and case managers as part of a feedback system designed to assess treatment progress as it unfolded in practice. The patients and therapists completed the Compass questionnaire, which included the MHI scales, typically preceding weekly sessions, but at intervals that varied across patients (mode = 7 days; $Mdn = 12.50$; $M = 14.57$, $SD = 9.48$). They did this 2–19 times each over the course of treatment. Patients who completed at least two questionnaires were included in our sample. The median number of sessions was 6; 67% of the patients had finished their treatment after 8 sessions and 87% after 16 sessions, and 5% of the patients had more than 25 sessions of treatment.

Data Analysis Strategy

The models included three levels: (a) sessions within patients, (b) patients within therapists, and (c) therapists. Data analyses were conducted with HLM 5 software (Raudenbush & Bryk, 2002; Raudenbush, Bryk, Cheong, & Congdon, 2001). We used an anchored model comparable to that used in the Elkin et al. (2006) study, as described in more depth in the Appendix. Anchoring treats patients as beginning from a common baseline to focus on change variance. Following previous work on dose–effect curves in psychotherapy (e.g., Hansen, Lambert, & Forman, 2002; Howard, Kopta, Krause, & Orlinsky, 1986), we modeled change in therapy as a negatively accelerating function of the number of sessions (but see Barkham et al., 2006). That is, the most rapid response was assumed to occur early in therapy. This curvilinear pattern is parsimoniously and conveniently approximated by a log-linear function of session number (e.g., Gibbons et al., 1993; Lambert, Hansen, & Finch, 2001; Lutz, Martinovich, Howard, & Leon, 2002).

At Level 1 (the session level), each patient's symptom intensity was modeled as a function of $\log_{10}$ of session number. We estimated two parameters: The intercept (the patient's MHI score at the first session) and the slope (the expected change in MHI score per $\log_{10}$ of session number). Session number was treated as a random variable; thus, there was no need to exclude cases from analyses just because they had different numbers of assessments or were missing data for particular sessions (see Nich & Carroll, 1997). The Level 1 intercept parameter was determined by the Session 1 MHI score; thus, the only variation was in the slopes, described as log-linear curves beginning at the observed baseline MHI (i.e., the intercept).

At Level 2 (the patient level), we searched for patient characteristics that predicted the Level 1 slope parameter. Our predictors at Level 2 included baseline levels of patients' symptom intensity on the MHI, the therapist-rated GAS, and three additional patient-rated variables: prior psychotherapy, chronicity, and treatment expectations. These predictors had shown an impact on change in a previous study (Lutz, Martinovich, & Howard, 1999). Including them allowed us to investigate and adjust for therapist differences in case mix.

At Level 3 (the therapist level), we modeled effects of therapists and effects of patient predictors (indicating interactions of therapists with patient characteristics); both were modeled as random effects. The full model thus partitioned MHI slope variance into three components: (a) session-to-session variance in MHI scores within patients; (b) patient variance in rates of change (slopes) within therapists, adjusted for patient-level predictors; and (c) therapist variance in their patients' mean MHI rate of change and therapist variance in effects of patient-specific predictors of slope variance.

We then calculated therapist effect percentages in two alternative ways. First, we calculated the percentage of overall variance explained by therapists by comparing therapist variance in slopes (Component 3) with total variance in MHI scores, including all three components (sessions within patients, patient slopes within therapists, and therapist variance in mean slopes). Second, we calculated the percentage of slope variance explained by therapists by comparing therapist variance with variance in the slopes, including only the last two components (patient slopes within therapists and therapist variance in mean slopes). We present both percentages for completeness, but we suggest that the second alternative, percentage of slope variance explained, more closely represents the conceptual meaning of therapist effects on outcome, insofar as the slopes represent each patient's improvement across treatment. Note that the first alternative (percentage of total variance) does not arise when outcome is assessed as symptom intensity at only one point (a posttreatment measure) or across two points (change scores or residual gain), as has been the case in most previous studies (see Elkin et al., 2006; Raudenbush & Bryk, 2002; Wampold & Brown, 2005).

Results

Therapist Effects

Table 1 shows the variance analysis for the three-level model after adjustment for fixed effects. Among the Level 2 patient

Table 1
Variance Decompositions and Explained Variance From Three-Level Analysis of Mental Health Index Scores

Random effect	Variance component	*df*	χ^2
Level 1 (assessments within patients)			
Temporal variation error	39.63		
Level 2 (patients within therapists; $N = 1{,}198$)			
Patient slopes	30.12	1015	2,099.31**
Level 3 (between therapists; $N = 60$)			
Therapist mean slope	6.90	59	146.95**
Global Assessment Scale	0.02	59	91.17**
Prior psychotherapy	0.21	59	81.53*
Result of variance–covariance components calculations among the Level 3 random effects			
T[a]	6.28		
Variance explained by patients (%)	39.62		
Variance explained by time (%)	52.12		
Variance explained by therapists (%)			
Overall	8.26		
Slope variance	17.25		

Note. Analysis controlled for patient severity and intake characteristics at Level 2.

[a] $T = \begin{bmatrix} 6.90 & 0.21 & -0.65 \\ 0.21 & 0.02 & 0.02 \\ -0.65 & 0.02 & 0.21 \end{bmatrix}$

* $p < .05$. ** $p < .01$.

intake characteristics, initial MHI, chronicity, and treatment expectation failed to yield significant random components at Level 3, so we omitted the random effects for these components from the final model, leaving a final model with three error components, as shown in Table 1.

As shown in Table 1, Level 1 variance was 39.63, and Level 2 variance was 30.12. Because we had three correlated random error components at Level 3, the Level 3 variance is the sum of the elements in the variance–covariance matrix, shown at the bottom of Table 1 (Nunnally & Bernstein, 1994; Raudenbush & Bryk, 2002). This yielded a Level 3 (therapist) variance of 6.28 and an overall variance of 76.04 (sum of Level 1, 2, and 3 variance estimates). Thus, 8.26% (6.28/76.04) of the overall variance in MHI scores was explained by differences among therapists. An additional 39.62% (30.13/76.04) of the variance was between patients (within therapists), and the remaining 52.10% (39.63/76.04) was explained by session-to-session variation within patients. If we consider only variance in patient log-linear slopes (36.40, the sum of Level 2 and 3 variance estimates), then therapist differences accounted for 17.25% (6.28/36.40) of the variance in rate of patient change.

A figure presenting the average change rates for the 10 therapists with the highest slopes and the 10 therapists with the lowest slopes, with controls for initial MHI, is available as an online supplement to this article. No therapist's average slope was negative, but there were substantial therapist differences in slopes. The average slope for the top therapist was 11.43 (expressed in MHI T score points per log of session number—i.e., more than one standard deviation change on the MHI during the first 10 sessions), as compared with an average slope of 2.82 for the bottom therapist (about a third of one standard deviation change during the first 10 sessions).

Table 2 shows the fixed effects of Level 2 predictors on patient rate of improvement, measured as the slope (rate of change) in MHI scores. All patient intake predictors were expressed in mean deviation form. The intercept (6.66) represents the mean rate of change of more than half a standard deviation on the MHI over the first 10 sessions. All five patient intake indexes had significant effects on MHI slopes. As Table 2 shows, higher MHI scores at Session 1, prior psychotherapy, and problem chronicity all predicted slower improvement (i.e., the coefficients were negative). Therapists' ratings on the GAS and patients' positive treatment expectations predicted more rapid improvement. Because the coefficients for fixed effects in Table 2 are in T score units (so $SD = 10$), dividing the effects of predictors by 10 yields an effect size statistic (analogous to Cohen's *d*). For example, the effect size associated with a 10-point (one standard deviation) difference in clinician GAS rating was 0.11 standard deviation units (a relatively small effect).

In a separate model that did not include therapist effects (not shown in the tables), the five intake predictors explained 43.7% of patient slope variance (a reduction in variance from 71.3 to 40.1). This percentage is somewhat higher than in previous analyses (cf. Lutz et al., 1999). Most of the variance was explained by initial MHI (40.2%); the other intake predictors added only 3.5%. Finding a large impact for initial MHI scores on subsequent MHI-measured improvement is consistent with previous reports on the impact of pretreatment characteristics on change rates (Lambert et al., 2001; Leon, Kopta, Lutz, & Howard, 1999; Lutz et al., 1999, 2005, 2006).

Cross-Validation

As a cross-validation, we randomly divided the therapists into two subsamples, each with 30 therapists, and repeated the three-level growth curve analyses separately on each subsample. The 30 therapists in Subsample 1 treated 578 patients, and the 30 therapists in Subsample 2 treated 620 patients. Table 3 shows the variance analysis for those two subsamples.

For Validation Subsample 1, the Level 1 variance was 37.84, and the Level 2 variance was 28.17. The Level 3 (therapist) variance—the sum of the elements in the variance–covariance matrix, shown at the bottom of Table 3—was 6.58, so the overall variance was 72.59 (sum of Level 1, 2, and 3 variance estimates).

Table 2
Fixed Effects on Rate of Change (Slope) of the Mental Health Index

Fixed effect for slope	Coefficient	*SE*	*t*
Intercept	6.66	0.447	14.89**
First Mental Health Index	−0.67	0.035	−19.44**
Global Assessment Scale	0.11	0.039	2.74**
Prior psychotherapy	−0.65	0.156	−4.14**
Chronicity	−0.43	0.162	−2.66**
Treatment expectations	0.69	0.256	2.69**

** $p < .01$.

Table 3
Variance Decompositions and Explained Variance From Three-Level Analysis of Mental Health Index Scores for the Two Cross-Validation Samples and the Extreme-Group-Excluded Sample

Random effect	Variance component		
	Validation Sample 1	Validation Sample 2	Extremes excluded
Level 1 (assessments within patients)			
Temporal variation error	37.84	41.23	37.71
Level 2 (patients within therapists)			
Patient slopes	28.17	29.43	31.93
Level 3 (between therapists)			
Therapist mean slope	6.99	6.08	2.72
Global Assessment Scale	0.01	0.03	0.02
Prior psychotherapy	0.04	0.75	0.18
Result of variance–covariance components calculations among the Level 3 random effects			
T	6.58	5.14	2.84
Variance explained by patients (%)	38.81	38.83	44.05
Variance explained by time (%)	52.13	54.39	52.03
Variance explained by therapists (%)			
Overall	9.06	6.78	3.92
Slope variance	18.94	14.87	8.17

Note. Analysis controlled for patient severity and intake characteristics at Level 2.

Thus, 9.06% (6.58/72.59) of the overall variance in MHI scores and 18.94% (6.58/34.75) of the variance in slopes was explained by differences between therapists in Subsample 1, whereas 38.81% (28.17/72.59) of the variance in MHI scores was explained by variance between patients (within therapists), and the remaining 52.13% (37.84/72.59) was explained by session-to-session variation within patients. For Validation Sample 2, the Level 1 variance was 41.23, the Level 2 variance was 29.43, and the Level 3 variance was 5.14. Thus, 6.78% (5.14/75.80) of the overall variance in MHI scores and 14.87% (5.14/34.57) of the variance in slopes was explained by differences between therapists, whereas 38.83% (29.43/75.80) of the variance in MHI scores was between patients (within therapists), and the remaining 54.39% (41.23/75.80) was explained by session-to-session variation within patients.

Extreme Therapists Excluded

As a further check, we excluded the 6 therapists (10%) with the most extreme scores (the 3 with the lowest and the 3 with the highest average change rates) and applied again the same modeling strategy (see Table 3). For this sample with extreme therapists excluded, the Level 1 variance was 37.71, the Level 2 variance was 31.93, and Level 3 therapist variance was reduced to 2.84. Thus, 3.92% (2.84/72.48) of the overall variance in MHI scores and 8.2% (2.84/34.77) of the variance in patient slopes was explained by differences between therapists in this sample, whereas 44.05% (31.93/72.48) of the variance in MHI scores was explained between patients (within therapists), and the remaining 52.03% (37.71/72.48) was explained by session-to-session variation within patients.

Discussion

In this large, real-world data set, our multilevel data-analytic approach showed that about 8% of the variance in the intensity of patients' symptoms, as measured before each session, and 17% of the variance in estimated rates of patient improvement (log-linear MHI slopes) was explained by therapist differences. Analyses included all three nested random factors inherent to psychotherapy evaluation designs (sessions, patients, and therapists) and controlled for patient baseline differences in five variables: patient- and therapist-rated symptom intensity (MHI and GAS, respectively), chronicity, previous treatment, and therapeutic expectancies. The proportion of the variance in rate of change attributable to therapist differences (i.e., 17%) was higher than that found by Wampold and Brown (2005) and Kim et al. (2006) as well as Crits-Christoph et al. (1991), who found that around 8% of the outcome variance could be explained by therapist effects. The split-half cross-validation analysis resulted in therapist effects of about 7%–9% of variance in MHI scores and 15%–19% of variance in rate of change explained by therapists. The analyses using a reduced data set with extreme therapists excluded also resulted in therapist effects of about 4% of total scores and 8% of slopes. This artificially reduced effect was significant and still sufficiently high to suggest that therapist effects are not due simply to occasional outliers, as suggested by Elkin et al. (2006).

The contention raised by Crits-Christoph et al. (2003) and Elkin et al. (2006) that therapist differences tend to be larger in naturalistic studies than in controlled trials finds some support in our findings. These data were acquired for administrative purposes instead for a formal clinical trial, and the therapist effect on outcome (understood in our study as rate of improvement) was approximately double that typically observed in the clinical trials. Naturalistic samples may include a wider range of therapist skill than do controlled clinical trials, in which therapists are selected and trained to conform to specified treatment protocols. Therapists considered as outliers in a controlled clinical trial might be considered merely as being at the ends of the distribution in a naturalistic sample. The best managed clinical trials may show the smallest therapist effects (e.g., Elkin et al., 2006). Estimates in the empirical literature of the amount of explained variance due to therapists in clinical trials versus naturalistic settings deserves further investigation, however. The relative influence of a number of possible contributing factors (e.g., the heterogeneity of patients and therapists, analytic methods, outcome measures used) has not been substantially investigated.

Our results show that a growth curve model that includes three levels of nesting does not necessarily eliminate therapist effects, as Elkin et al. (2006) suggested. Previous studies that used only two-point (pre- and posttreatment) assessments (change scores or residual gain scores) might have found smaller therapist effects because their assessments were more vulnerable to random variation over time than were the slopes we modeled at Level 1, which were based on a median of six sessions per patient. That is, the earlier lower estimates could merely reflect less reliable outcome measurement. This possibility as well as others, such as variation

in severity of patient disturbance in different samples, should be also explored in future studies about therapist effects.

The strength of this study is also its primary limitation. It was conducted with a large database of patients and therapists in a naturalistic setting. Treatments were conducted under natural conditions rather than under the monitored conditions of a randomized controlled trial. Thus, our results represent therapy in the way it is delivered rather than the way it perhaps should be delivered. The study's relatively greater realism and external validity have come at the cost of a lack of experimental control and lack of detailed information about many aspects of the participants and the treatments.

The naturalistic nature of our data also makes them subject to potential confounds. In principle, the therapist effects might have been overestimated because of the lack of random assignment of patients to therapists. For example, some of the therapists might have specialized in specific patient groups, or some types of patients might have had different insurance coverage, influencing their representation in therapists' caseloads. Our controlling for the effects of some baseline patient attributes can be seen as a strength, but many other variables may influence patient response to treatment, such as verbal ability, relational orientation, and social support (Garfield, 1994; Lambert & Bergin, 1994), and these additional predictors are potential confounds. If therapists' caseloads differ systematically in one or more of these predictors, these patient differences might have artificially inflated the therapist variance estimates reported in this article.

Conversely, if differential assignment to therapists was done responsively (Stiles, Honos-Webb, & Surko, 1998), this might have reduced therapist effects in comparison with random assignment. That is, if each therapist was assigned the types of patients he or she was most capable of treating, then all patients might have tended to receive something closer to optimum treatment. One could reduce variability by avoiding the poor average outcomes that might otherwise accrue to therapists whose talents or interests focus on a narrow range of types of patients. Conversely, random assignment of patients to a naturalistic sample of therapists might lead to an even larger variation in therapist mean outcome than we observed.

In the present work, we have found substantial therapist effects after controlling for a limited array of patient intake predictors. Future work might consider differential effectiveness on the basis of profiles of therapist personal or professional or patient characteristics. The hypothesis that therapists are differentially effective with particular types of patients has not received much support in the empirical literature (Elkin et al., 2006; Luborsky et al., 1986; Pilkonis, Imber, Lewis, & Rubinksy, 1984; Shapiro et al., 1989). However, the previous investigations were not completed on large naturalistic databases with repeated-measurement designs and new available methodological tools. Results showing how therapists achieve successful treatment effects in different ways might be used to improve clinical training as well as supervision and have the potential to improve case assignment in routine care.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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,* 160–167.

Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., Carroll, K., Perry, K., et al. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1,* 81–91.

Crits-Christoph, P., & Gallop, R. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. *Psychotherapy Research, 16,* 178–181.

Crits-Christoph, P., & Mintz, J. (1991). Implication of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology, 59,* 20–26.

Crits-Christoph, P., Tu, X., & Gallop, R. (2003). Therapists as fixed versus random effects—Some statistical and conceptional issues: A comment on Siemer and Joormann. *Psychological Methods, 8,* 518–523.

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: Disentangling therapist from therapies. *Clinical Psychology: Science and Practice, 6,* 10–32.

Elkin, I. (2006). Rejoinder to commentaries by Stephen Soldz and Paul Crits-Christoph on therapist effects. *Psychotherapy Research, 16,* 182–183.

Elkin, I., Falconnier, L., Martinovich, Z., & Mahoney, C. (2006). Therapist effects in the NIMH Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16,* 144–160.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., et al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–982.

Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The Global Assessment Scale: Procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry, 33,* 766–771.

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 190–228). New York: Wiley.

Garfield, S. L. (1997). The therapist as a neglected variable in psychotherapy research. *Clinical Psychology: Science and Practice, 4,* 40–43.

Gibbons, R. D., Hedeker, D., Elkin, I., Waterneaux, C., Kraemer, H., Greenhouse, J. B., et al. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry, 50,* 739–750.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9,* 329–343.

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

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). The evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51,* 1059–1064.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., & Shear, M. K. (2001). Therapists, therapist variables, and cognitive–behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 65,* 747–755.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Kim, D.-M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random effects modeling of the NIMH TDCRP data. *Psychotherapy Research, 16,* 161–172.

Lambert, M. J., & Bergin, A. E., (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 143–189). New York: Wiley.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused

research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159–172.

Lambert, M. J., & Okiishi, J. C. (1997). The effects of the individual psychotherapist and implications for future research. *Clinical Psychology: Science and Practice, 4,* 66–75.

Leon, S. C., Kopta, S. M., Lutz, W., & Howard, K. I. (1999). Predicting patients' responses to psychotherapy: Are some more predictable than others? *Journal of Consulting and Clinical Psychology, 67,* 698–704.

Luborsky, L., Crits-Christoph, P., McLellan, A. T., Woody, G., Piper, W., Liberman, B., et al. (1986). Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry, 56,* 501–512.

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E., & Grissom, G. (2001). Assessing treatment progress with individualized models of predicted response. *Journal of Consulting and Clinical Psychology, 69,* 150–158.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., et al. (2005). Predicting rate and shape of change for individual clients receiving psychological therapy: Using growth curve modeling and nearest neighbor technologies. *Journal of Consulting and Clinical Psychology, 73,* 904–913.

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology, 67,* 571–577.

Lutz, W., Martinovich, Z., Howard, K. I., & Leon, S. C. (2002). Outcomes management, expected treatment response and severity adjusted provider profiling in outpatient psychotherapy. *Journal of Clinical Psychology, 58,* 1291–1304.

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., et al. (2006). Empirically and clinically useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment, 18,* 133–141.

Lyons, J. S., Howard, K. I., O'Mahoney, M. T., & Lish, J. D. (1997). *The measurement & management of clinical outcomes in mental health.* New York: Wiley.

Martinovich, Z., & Helgerson, J. (in press). Applications of trajectory analysis in research and outcomes management. In J. Lyons & D. Aron-Weiner (Eds.), *Behavioral health care strategies: Total clinical outcomes management.* Kingston, NJ: Civic Research Institute, Inc.

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology, 71,* 404–409.

Nich, C., & Carroll, K. (1997). Now you see it, now you don't: A comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *Journal of Consulting and Clinical Psychology, 65,* 252–261.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

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

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

Pilkonis, P. A., Imber, S. D., Lewis, P., & Rubinsky, P. (1984). A comparative outcome study of individual, group, and conjoint psychotherapy. *Archives of General Psychiatry, 41,* 431–437.

Project MATCH Research Group. (1998). Therapist effects in three treatments for alcohol problems. *Psychotherapy Research, 8,* 455–474.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2001). *HLM 5: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Ricks, D. F. (1974). Supershrink: Methods of a therapist judged successful on the basis of adult outcomes of adolescent patients. In D. F. Ricks, M. Roff, & A. Thomas (Eds.), *Life history research in psychopathology* (pp. 288–308). Minneapolis: University of Minnesota Press.

Serlin, R. C. Wampold, B. E., & Levin, J. R. (2003). Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: A comment on Siemer and Joormann (2003). *Psychological Methods, 8,* 524–534.

Shapiro, D. A., Firth-Cozens, J., & Stiles, W. B. (1989). The question of therapists' differential effectiveness: A Sheffield Psychotherapy Project addendum. *British Journal of Psychiatry, 154,* 383–385.

Siemer, M., & Joormann, J. (2003). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods, 8,* 435–544.

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

Stiles, W. B., Honos-Webb, L., & Surko, M. (1998). Responsiveness in psychotherapy. *Clinical Psychology: Science and Practice, 5,* 439–458.

Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Brody, P. E. (2003). Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. *Journal of Consulting and Clinical Psychology, 71,* 963–972.

Vessey, J. T., & Howard, K. I. (1993). Who seeks psychotherapy? *Psychotherapy, 30,* 546–553.

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research, 16,* 184–187.

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73,* 914–923.

Appendix

The Three-Level Growth Curve Model

Assuming an underlying log-linear course of recovery, each patient's outcome score may be modeled (Level 1) as a function of session number, as follows:

$$\text{Outcome}_{spt} = \beta_{0pt} + \beta_{1pt}\text{Log(Session)}_{spt} + e_{spt}. \quad \text{(A1)}$$

Outcome_{spt} is the observed outcome score at a particular session (s) for a patient (p) treated by a specific therapist (t). The β_{0pt} parameter (intercept) is a patient's expected outcome score at the first session. The β_{1pt} parameter (slope) is the expected change in outcome scores per $\log_{10}$ of session number. Session number is a random variable, with patients differing on the number and spread of sessions with assessments. The random error term, e_{spt}, refers to normally distributed deviations from expected values for patient i at session t. This model is referred to as a Level 1 model.

At Level 2 (the patient level), variation in Level 1 intercepts and slopes is modeled as follows:

$$\beta_{0pt} = 0 + 1 \cdot \text{Outcome}_{0pt}$$

$$\beta_{1pt} = \pi_{10t} + \pi_{11t}\text{Outcome}_{1pt} + \pi_{12t}\text{GAS}_{1pt} + \pi_{13t}\text{Prior Psychotherapy}_{1pt} + \pi_{14t}\text{Chronicity}_{1pt} + \pi_{15t}\text{Treatment Expectations}_{1pt} + r_{1pt}. \quad \text{(A2)}$$

Intercepts and slopes at Level 2 are modeled as linear functions of patient and therapist variation in baseline scores. By fixing π_{00t} and π_{01t} terms at 0 and 1, respectively, we force the model to predict each patient's baseline status at Session 0, and there is no patient-specific error term for baseline levels. At the patient level, the error term is specific to patient (within therapist), and the fixed effect coefficients are specific only to therapist. This anchoring procedure (each patient's change trajectory passes through his or her baseline score) follows the same data-analytic strategy as described in Elkin et al. (2006) and Martinovich and Helgerson (in press). It is analogous to an analysis of covariance procedure commonly used in clinical trial analyses (including pretest as the covariate and measuring adjusted slopes instead of adjusted outcomes on the posttest variable). Because anchoring eliminates the need for a random component for the intercept, all change variance is allocated to the slope term. In addition to simplifying the model, anchoring enhances slope reliability (all reliable change variance is allocated to the slope random component).

It is possible to search for predictors of intercept and slope parameters by constructing Level 2 models in which Level 1 coefficients (only slopes in this example) are dependent variables. The random effect in these models refers to the reliability of unexplained variability, whereas the fixed effect refers to the effect of factors influencing average slope or intercept. The dispersion of random effects is represented by a variance–covariance matrix; therefore, error components may covary and have unequal variances.

Initially, the model only included a single predictor at Level 2 (baseline outcome variable). We subsequently augmented the model by including estimates of patients' level of presenting variables to adjust for differences in case mix for the therapist. Therefore, we added five significant predictors that have shown their impact on individual differences in change in previous studies (Lutz et al., 1999).

At Level 3 (the therapist level), variation in patient-level coefficients is broken into fixed and random components, as follows:

$$\pi_{10t} = \gamma_{100} + U_{10t}$$

$$\pi_{11t} = \gamma_{110}$$

$$\pi_{12t} = \gamma_{120} + U_{12t}$$

$$\pi_{13t} = \gamma_{130} + U_{13t}$$

$$\pi_{14t} = \gamma_{140}$$

$$\pi_{15t} = \gamma_{150}.$$

Because including an error component for π_{11}, π_{14}, and π_{15} did not result in a significant variance in components, these terms were excluded, leaving a final model with three error components estimating (a) therapist variance for their average rates of change (averaged across their patients), adjusted for baseline predictors (with the Level 3 variance and covariance components added together for U_{10t}, U_{12t}, and U_{13t}; see matrix in Table 1); (b) patient variance (within therapist) in rates of change, adjusted for baseline status (r_{pt}); and (c) session-specific variance of outcomes away from estimated values on the basis of each patient's anchored linear trajectory (e_{spt}).

Received August 25, 2005
Revision received August 22, 2006
Accepted August 28, 2006 ■