Journal of Consulting and Clinical Psychology
2005, Vol. 73, No. 5, 914–923

0022-006X/05/$12.00 DOI: 10.1037/0022-006X.73.5.914

Estimating Variability in Outcomes Attributable to Therapists: A Naturalistic Study of Outcomes in Managed Care

Bruce E. Wampold
University of Wisconsin—Madison

George S. (Jeb) Brown
Center for Clinical Informatics

To estimate the variability in outcomes attributable to therapists in clinical practice, the authors analyzed the outcomes of 6,146 patients seen by approximately 581 therapists in the context of managed care. For this analysis, the authors used multilevel statistical procedures, in which therapists were treated as a random factor. When the initial level of severity was taken into account, about 5% of the variation in outcomes was due to therapists. Patient age, gender, and diagnosis as well as therapist age, gender, experience, and professional degree accounted for little of the variability in outcomes among therapists. Whether or not patients were receiving psychotropic medication concurrently with psychotherapy did affect therapist variability. However, the patients of the more effective therapists received more benefit from medication than did the patients of less effective therapists.

Keywords: therapist variability, psychotherapy, outcomes, multilevel analyses, managed care

During the last 2 decades, data from clinical trials have been reanalyzed with the goal of estimating the proportion of variability in outcomes that is attributable to therapists (e.g., Blatt, Sanislow, Zuroff, & Pilkonis, 1996; Crits-Christoph et al., 1991; Crits-Christoph & Mintz, 1991; Huppert et al., 2001; Kim, Wampold, & Bolt, in press). Synthesizing the resultant estimates of variability in outcomes attributable to therapists has been difficult because of (a) significant variability among estimates, (b) factors that affect the size of therapist effects, and (c) inconsistencies in the manner in which therapist effects are conceptualized and calculated.

With regard to the range of estimates, Crits-Christoph and Mintz (1991) reanalyzed data from 10 clinical trials and found that the proportion of variance due to therapists ranged from 0% to 13.5%, on the basis of the mean of variables within studies. Crits-Christoph and Mintz (1991), on the basis of results from 27 treatment groups, found 8.6% of the overall variance in outcomes was attributable to therapists, but the range was from 0% to nearly 50% (and up to 73% for individual variables). In a reanalysis of the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP) data and using various multilevel modeling methods, Kim et al. (in press) found that about 8% of the variance in the outcomes of psychotherapy conditions was due to therapists, but that this percentage varied by outcome measure analyzed and the manner in which therapist variance was modeled. Huppert et al. (2001), in a reanalysis of cognitive–behavioral treatment in a multicenter trial for panic disorder, reported therapist effects sizes for the various measures ranging from 1% to 18%.

With regard to factors that account for this variability in estimates, Crits-Christoph et al. (1991) were interested in the characteristics of the trials that could be related to the amount of therapist variability observed, including type of treatment (dynamic or behavioral), whether a manual was used, the experience of the therapist, and the length of treatment. When considered as a set of variables, the use of a manual and a greater level of experience were associated with less therapist variability. As well, Crits-Christoph et al. (1991) reported that more recently conducted trials had less variability than did older ones. However, in an analysis of a rigorous and recently conducted trial, Huppert et al. (2001) reported variability that seems equal to or in excess of the average value reported in earlier studies. These estimates indicate that variability in outcomes attributable to therapists is an important factor, as the proportion of variance due to the type of treatment delivered is at most 1% or 2%, and the variability due to alliance, the most prominent common factor, is around 5% (Wampold, 2001).

Another problem with comparing the results of studies is that various statistical models have been used to determine therapist effects, with each model conceptualizing therapist variance in a different manner. Early in the evolution of estimating these effects, Crits-Christoph and Mintz (1991) discussed the methodological issues that arise in the analysis of therapist effects. A critical decision is whether to treat therapists as a fixed or a random factor (see Crits-Christoph & Mintz, 1991; Elkin, 1999; Serlin et al., 2003; Siemer & Joormann, 2003; Wampold & Serlin, 2000, for a discussion of fixed and random models in psychotherapy research). If therapists are treated as fixed, the results are conditioned on the particular therapists included in the clinical trial. Although restricting the generality of the results yields an increase

Bruce E. Wampold, Department of Counseling Psychology, University of Wisconsin—Madison; George S. (Jeb) Brown, Center for Clinical Informatics, Salt Lake City, Utah.

Bruce E. Wampold has consulted with PacifiCare Behavioral Health (PBH) and George S. (Jeb) Brown has a continuing consulting arrangement with PBH. All analyses and preparation of this article were conducted independently and no editorial oversight was exercised by PBH, as per an agreement between PBH and the authors prior to undertaking this project.

Correspondence concerning this article should be addressed to Bruce E. Wampold, Department of Counseling Psychology, 321 Education Building—1000 Bascom Mall, University of Wisconsin, Madison, WI 53562. E-mail: wampold@education.wisc.edu

in power to test main effects, conclusions about a particular small set of therapists would appear to be an unreasonable restriction (see Serlin et al., 2003; Siemer & Joormann, 2003).

More informative results are obtained when therapists are considered as being randomly selected from a population of therapists so that conclusions can be made about therapists in general (or, in the absence of random selection, about therapists similar to the ones used in the trial; Serlin et al., 2003). The differences in the models are summarized by Siemer and Joormann (2003):

> The crucial question is whether it is justified to treat providers as a random effect thereby seeking to generalize to a population of providers or whether one should treat providers as a fixed effect thereby restricting the inference to the providers included in that particular study, that is, to make statistical inference conditional on the set of providers included in the study. (p. 500)

Huppert et al. (2001) entered therapists in an ordinary least squares analysis, thus treating them as a fixed factor. Blatt et al. (1996) segregated therapists into classes on the basis of their outcomes and used analyses of variance (ANOVAs) to examine group differences, again treating them as a fixed effect. Kim, Wampold, and Bolt (in press), on the other hand, used multilevel analysis (Snijders & Bosker, 1999) to treat therapists as a random factor.

Another methodological issue is whether the effect due to therapists reported (in either the fixed or random model) is a sample value or an estimate of a population parameter. When fixed effects are used, R^2 in the regression context or η^2 in the ANOVA context are sample values that tend to overestimate the true proportion of variance due to therapists. This is especially problematic in the context of clinical trials because typically there are few patients per therapist, increasing the magnitude of the bias. In the fixed effects case, shrunken R^2 and ω^2 correct for this bias and should be reported. In the random effects context, the appropriate estimate is the intraclass correlation coefficient ρ_I, which indexes the covariation of scores within therapists to the total variation among scores (see Wampold & Serlin, 2000).

In spite of the problems with finding an average value to assign to variability in outcomes attributable to therapists, it appears that when therapists are treated as random and the appropriate statistical models are used, about 8% of the variability in outcomes can be attributed to them (see Kim et al., in press). This is a result that is generalizable to the type of therapist used in clinical trials, in which therapists typically are selected for their skill, are especially trained, receive supervision, and are guided by a manual (Elkin, 1999; Wampold & Serlin, 2000; Westen, Novotny, & Thompson-Brenner, 2004). Therapist effects determined in clinical trials are also restricted by other contextual variables, such as the homogeneity of the patients enrolled in these trials (Westen et al., 2004). To our knowledge, no estimate of the variability in outcomes attributable to therapists in clinical practice has been reported. Therapist variability in clinical settings has implications for quality assurance and the management of care because such variability indicates that certain therapists are consistently producing below-average outcomes.

The first purpose of the present study was to estimate the proportion of variability of outcomes attributable to therapists in independent practice. Conjecture and the results of Crits-Christoph et al.'s (1991) analysis would suggest that therapist variability in practice would far exceed that found in clinical trials because the therapists are delivering a variety of treatments to heterogeneous patients, without supervision or training or the guidance of a manual (see also Westen et al., 2004). However, one has to be cautious about this prediction for two reasons. First, the effects of treatment, manuals, supervision, and training have not been shown to be robust predictors of outcomes generally (see Wampold, 2001). Second, the heterogeneity of patients in practice implies that the total variation, which forms the denominator of ratios indexing therapist effects, would be greater, thereby decreasing such effects.

Given that, in most instances, variability among therapists has been detected, it is important to identify therapist variables (characteristics and actions) that are associated with this variability. Although therapist variables have been studied for decades, few have been reliably shown to be related to variability in outcomes that is attributable to therapists (Beutler et al., 2004). In the analysis of clinical trial data, many therapist variables have been found to be unrelated to outcomes. For example, Blatt et al. (1996), in the reanalysis of the three active treatments (viz., imipramine plus clinical management; cognitive–behavioral therapy [CBT]; and interpersonal psychotherapy) of the NIMH TDCRP, found that therapist age, gender, race, religion, marital status, clinical experience (both generally and with long- and short-term dynamic therapy, CBT, behavior therapy, and eclectic therapy) were not related to therapist effectiveness. However, therapist effectiveness was positively related to the use of psychological as opposed to biological interventions, psychological mindedness, and expected length of treatment. Huppert et al. (2001), in the analysis of CBT therapists treating panic disorder, found that effectiveness of therapists was not related to their age, gender, gender match, and experience with CBT, although on some variables, it was related to overall experience in conducting psychotherapy. A secondary purpose of the present study was to examine therapist and patient variables that might be associated with therapist effectiveness.

Method

The data analyzed in the present study were outcomes of patients seen by providers of PacifiCare Behavioral Health (PBH), a managed care organization. PBH has instituted a system to assess outcomes to increase the benefits attained by their patients (Brown, Burlingame, Lambert et al., 2001; Matumoto, Jones, & Brown, 2003). This section describes the outcome measure and administrative data available for analysis, the participants (patients and therapists), and the procedures used to prepare the data for analysis.

Outcome Measure

PBH used a 30-item, self-report questionnaire that was derived from the Outcome Questionnaire (OQ-45; Lambert, Gregersen, & Burlingame, 2004). PBH desired a shorter version of the OQ-45 that clinicians could easily administer as a paper-and-pencil instrument, that they could score prior to the treatment session, and that also retained the desirable psychometric properties of the longer questionnaire. Consequently, PBH contracted with Lambert and Burlingame to adapt the OQ-45 to the 30-item version, which is referred to as the Life Status Questionnaire (LSQ), a proprietary label for PBH.

The 30 items for the LSQ were selected from the OQ-45 item pool on the basis of sensitivity to change as estimated from a large-scale study of patients undergoing treatment in a variety of settings (Lambert, Hatfield, et al., 2001; Vermeersch, Lambert, & Burlingame, 2000). The LSQ measures

three aspects of functioning: (a) subjective discomfort, (b) interpersonal relationships, and (c) social role performance. Moreover, the LSQ contains items addressing problems common to a wide variety of disorders and reflecting quality of life (Lambert, Hatfield, et al., 2001). The scores obtained from a large sample of patients treated in an independent practice setting yielded a coefficient alpha of .94 and a test–retest reliability of .80 after a 3-week interval between administrations (Lambert, Hatfield, et al., 2001). Validity was established by an expected pattern of correlations with measures of mental health status including the Symptom Checklist 90R ($r = .70$; Derogatis, 1977), Beck Depression Inventory ($r = .61$; Beck, Ward, Mendelson, & Erbaugh, 1961), Inventory of Interpersonal Problems ($r = .62$; Horowitz, Rosenberg, Baer, Ureno, & Villasensor, 1988), and Social Adjustment Scale ($r = .59$; Weissman & Bothwell, 1976). The LSQ also adequately differentiated clinical samples from community samples and was sensitive to change (Lambert, Hatfield, et al., 2001).

Data Collection

PBH uses the LSQ across its network of providers; individual use is voluntary on the part of the patients and clinicians. At the time services are authorized, PBH mails a packet of LSQ forms to the therapist, with instructions to administer the questionnaire at the first, third, and fifth sessions, and at every fifth session thereafter. The clinician forwards the information to PBH by faxing the form to the toll free number provided for that purpose.

The number of clinicians and the percentage of patients using the questionnaires have increased year by year. During the first quarter of 2001, 33% of adults receiving psychotherapy services completed at least one LSQ during their episode of care. By the first quarter of 2004, 70% of patients receiving psychotherapy services completed at least one LSQ and over 60% of these completed multiple assessments during the course of the treatment episode.

Participants

The sample for this study consisted of adults (aged 18 or older) who began outpatient psychotherapy between January 1, 2001, and December 31, 2002. All patients included in this study had the opportunity to complete at least 6 months of treatment (i.e., data analyzed included all LSQs completed up to June 30, 2003).

The only patient variables available for analysis were gender, age, diagnosis, and, in some cases, psychotropic medication status. PBH, like all managed care companies, does not routinely capture demographic data such as patient race, education, or income level. For providers in independent practice, the variables available included age, gender, professional degree (or license type), and years of practice.

Defining Episodes of Care and Merging Data Sets

Two sources of administrative data were important to this study: claims submitted for outpatient psychotherapy services and those submitted for medications. It was necessary to merge data from these two administrative databases with the outcome data (i.e., LSQ scores) to get an accurate picture of the number of psychotherapy sessions provided to the patients and whether the patient took a psychiatric medication concurrent with psychotherapy.

In outpatient mental health services, patients receive services irregularly. Often patients see a provider on regular basis, then stop out for several weeks for various reasons and return to the provider at a later time. Because data were not available with regard to either therapist or patient-reported termination, it was necessary to define an episode of care arbitrarily on the basis of the continuity of care. For our purposes, an episode of care was defined as individual psychotherapy services from a single provider without an interruption of more than 90 days.

The voluntary and largely unmonitored process for administration of the LSQ inevitably resulted in missing data. To ensure that a case had sufficient LSQ data to make sense of patient progress, we used various rules to include or exclude cases. Each case had to have at least two LSQs, the first of which must have been labeled as Session 1 or Session 2, and no two administrations that were more than 90 days apart. Claims data were similarly organized into episodes of care, matching the patient with a single provider without an interruption of more than 90 days between services.

Patients who began treatment in 2001 had the possibility of data collected over a longer period of time than those who began treatment in 2002. For this reason the period for an episode was limited to the first 180 days of treatment to assure comparability of results from 2001 and 2002. The data used in this study consisted of the first and last LSQ available for each episode of care. Only therapists with four or more patients in the data set were included for this analysis.

The PBH data contained information with regard to the efficacy of psychotherapy in the presence or absence of psychotropic medications for a limited subset of patients. Pharmacy data were available only for patients treated during 2002 and then only if the pharmacy benefit was covered by a PBH-affiliated plan. In many instances the pharmacy benefit was provided through a nonaffiliated plan, making the data unavailable for purposes of this study.

If pharmacy data were available, we created separate episodes for each drug prescribed. A drug episode was considered to have continued as long as the prescription was refilled without interruption. If the patient did not refill the prescription before the last filling had been exhausted, then the end date of the episode was the date the patient was expected to run out of the medication. Once we had data organized into drug episodes, we merged the data matching the start dates of the episodes of care from the three sources of data: outcome questionnaires, service claims, and pharmacy claims. We began by merging the claims data episodes with the LSQ episodes. If we found that the clinician had been treating the patient more than 21 days prior to the first LSQ record, we excluded the case. Likewise, if the clinician submitting the LSQ data had not provided the majority of services to that patient the case was excluded from the final data set. Because of the time lag for claims to be submitted and processed, some of the cases starting treatment in the second half of 2002 did not have complete claims data for the period of the LSQ episode; cases were excluded if the last LSQ in the episode occurred later than the last claims record available for that episode.

To investigate the impact of medication in addition to psychotherapy, we needed to identify patients who either (a) did not receive any medication or (b) began a psychotropic medication concurrent with the start of psychotherapy. Patients who were receiving psychotropic medications prior to psychotherapy created ambiguity because we did not know the response to the medication prior to therapy. Consequently, the analyses conducted in this study with regard to medication were restricted to patients who had the PBH pharmacy benefit (i.e., determination could be made confidently whether or not they were prescribed a psychotropic medication) and who either had no psychotropic medication claims or who had had psychotropic medication claims within two weeks of the onset of psychotherapy.

Using these rules, we identified a sample of 581 therapists treating 6,146 patients. For the therapists, 72.3% were female and 27.7% were men, the mean age was 51.5 years ($SD = 14.87$; $Mdn = 53.0$), mean number of years of experience was 21.2 ($SD = 7.65$; $Mdn = 21.0$), and mean number of patients seen in the data analyzed was 9.68 ($SD = 5.61$; $Mdn = 8.0$). With regard to degree of the provider, 30.3% had doctoral degrees (PhD, EdD, or PsyD), 63.7% had master's degrees (master's in counseling, marriage and family therapy, social work, or related field), 3.6% had medical degrees, and 2.4% had other or unknown degrees. The primary diagnoses of the patients were depression disorders (46.3%), adjustment disorders (30.2%), and anxiety disorders (11.0%); see Table 1 for more information. No information was available with regard to comorbidity but we assumed that the patients exhibited the degree of comorbidity typical of

Table 1
Residualized Scores by Diagnostic Group

Diagnostic group	*n*	*M*	*SD*
Adjustment	1,854	−0.57	12.09
Anxiety	675	−0.55	12.56
Bipolar	193	1.83	14.41
Depression	2,848	0.47	12.90
Posttraumatic stress disorder	136	1.28	12.26
Other	299	−0.20	12.26

Note. $F(5, 5999) = 2.86$, $p = .014$.

community samples. For the patients, 72.3% were women and 27.7% were men, the mean age was 39.8 ($SD = 10.80$, $Mdn = 40.0$), and the mean number of sessions was 10.63 ($SD = 8.08$, $Mdn = 8.0$). Of the subgroup of patients for which pharmacy data were available, 1,083 were classified as receiving psychotherapy only, whereas 586 met the criteria for psychotherapy plus medications. The remainder of the sample either did not have pharmacy data available or if it was available, the start date of the medication(s) did not coincide with the start date of the psychotherapy.

Analysis and Results

Although the data structures appear quite simple (i.e., patients nested within therapists), the statistical models must be correctly specified to obtain appropriate estimates of therapist variability. In these analyses, therapists were considered a random factor so that conclusions could be made about therapists in general. In modeling the variability in outcomes due to therapists, the relationship between the initial level of severity and final outcome must be considered. In this data set, the last LSQ was highly correlated with the first LSQ ($r = .69$), a result that is consistent with intervention studies in general (pretest typically is highly correlated with posttest scores). Methods used need to take into account that patients were not randomly assigned to therapists, so that therapists differed in the severity of the patients they saw in this data set. Hence, it was necessary to take into account the first–last LSQ correlation.

When a two-level analysis is conducted (patient and therapist), one must take care in specifying how the first and last LSQ are related because two such relationships must be considered. First, one can examine the regression of the last LSQ onto the first LSQ within therapists. That is, for a particular therapist, there is a relationship between the scores before and at the end of treatment. These regressions then may be pooled across the therapists, assuming that the relationship is constant across them (see Figure 1, left panel). Or alternatively the regression slopes among therapists could be allowed to vary (see Figure 1, right panel). Second, there is a between-therapist regression. Some patients are initially more distressed than others; presumably across the therapists, the mean LSQ for a therapist on the first administration is correlated with the mean LSQ at the end of treatment. The multilevel analyses here will account for within-therapist and between-therapist regressions. It should be noted that the common practice of computing residualized gain scores will be flawed if the within-group and between-group regressions vary significantly. The various models are presented hierarchically in the sections below.

Unconditional Model

The first model considered only the variance components related to therapists and patients using the last LSQ. This analysis was the typical variance component or random-effects model ANOVA and yielded the estimates necessary to determine the proportion of variability due to therapists. The estimate of therapist variance, denoted by $\hat{\sigma}^2_{\text{ther}}$, was equal to 24.53 ($SE = 3.27$) and the estimate of the patient or error variance, denoted by $\hat{\sigma}_\varepsilon$, was equal to 288.59 ($SE = 5.47$). The proportion of variance due to therapists

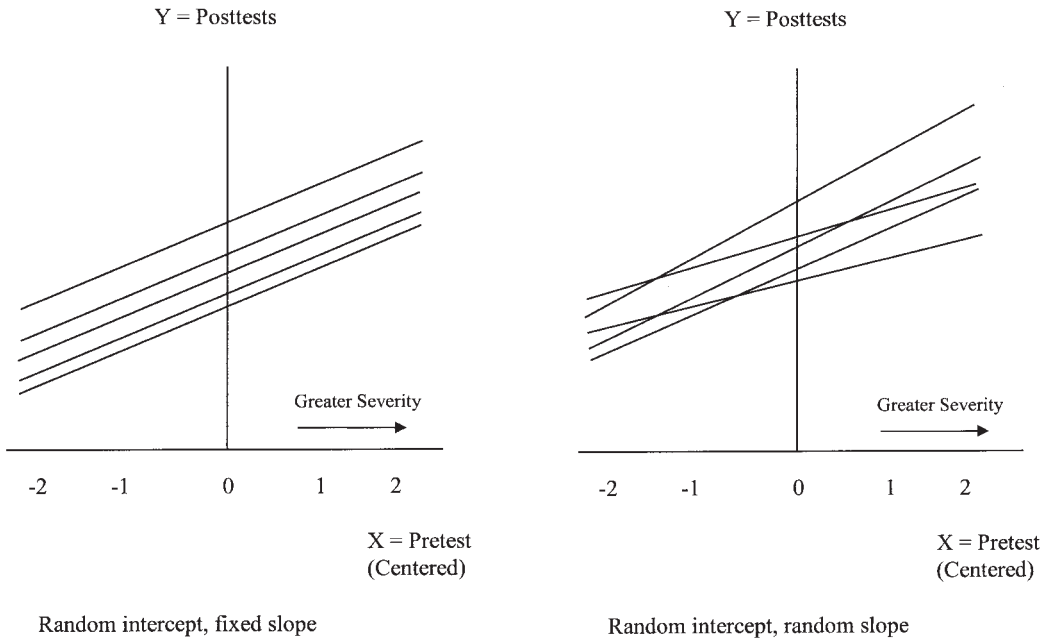

Figure 1. Hypothetical therapist slopes.

is given by the intraclass correlation coefficient, defined in the following way (Snijders & Bosker, 1999; Wampold & Serlin, 2000):

$$\hat{\rho}_I = \frac{\hat{\sigma}^2_{\text{ther}}}{\hat{\sigma}^2_{\text{ther}} + \hat{\sigma}^2_{\varepsilon}} = \frac{24.53}{24.53 + 288.59} = .078.$$

Essentially, this coefficient is the estimate of the population proportion of variance due to therapists divided by the total variance. Thus, in the present sample, about 8% of the variance in outcomes of the PBH patients was due to the therapists. This is a sizable proportion of variance and, given the standard error, is significantly larger than zero ($p < .0001$).

Models Conditional on Initial LSQ

Much of the variability in outcomes was due to the initial level of severity and therefore the next step in the analysis was to account for the first LSQ by conducting an analysis involving two levels—patient level, often referred to as *Level 1*, and therapist level, often referred to as *Level 2* (Snijders & Bosker, 1999; Raudenbush & Bryk, 2002). In the model tested, we considered both within-therapist and between-therapist regressions and allowed the slopes within therapists to vary across therapists.[1] A decision had to be made relative to parameterization of the initial LSQ, as this variable can be centered at the grand mean or centered at the mean for each therapist. We conducted the analysis using both parameterizations, which yielded the finding that the within-therapist regression coefficient was not too different from the between-therapist regression coefficient. Although parameterization does not affect variance estimates, it does have important implications for the use of residualized change scores (see below).

For our purposes, the variance estimates are central to understanding therapist effects; these estimates are presented in Table 2. Four variance components were estimated: therapist intercept, therapist slope, the covariance of intercept and slope, and unexplained patient variance (i.e., error). All of these estimates were made taking into account differences found in the initial LSQ, both within therapists and among therapists. The therapist intercept variance, τ^2_0, refers to the variance of the mean last LSQ for each therapist (again, having already taken into account the first LSQ). This estimate was relatively large (viz., 8.469) and significantly greater than zero. Therapist slope variance τ^2_1 is the variance in slopes among therapists when the last LSQ is regressed onto the first LSQ; this variance was small (viz., 0.142), although statistically different from zero. That is, the relationship between the first and last LSQ is not constant across therapists, although the differences were small. The intercept–slope covariance, τ^2_{01}, which was also small (viz., 0.016) but statistically significant, indicates that slope and intercept are correlated (in this case the variability among therapists is greater when the initial severity is greater). Finally, the patient variance, σ^2_{ε}, estimates the unexplained variance at Level 1.

Table 2
Estimates of Variance of Random Effects With First Life Status Questionnaire

Parameter	Description	Estimate	*SE*	*p*
τ^2_0	Therapist intercept variance	8.469	1.371	<.0001
τ^2_1	Therapist slope variance	0.142	0.053	.007
τ^2_{01}	Intercept–slope covariance	0.016	0.004	<.0001
σ^2_{ε}	Patient variance (error)	149.76	2.942	<.0001

The important determination here is the proportion of variability in outcomes due to therapists, taking into account the first LSQs. Because the slopes are allowed to vary, the proportion of variance due to therapists is a function of the initial level of severity (see Figure 1, where the variation among the lines fluctuates across the *x*-axis). In this case, we provide the proportion at the mean level of initial severity (see Kim et al., in press; Snijders & Bosker, 1999), which is given by

$$\frac{\hat{\tau}^2_0}{\hat{\tau}^2_0 + \hat{\sigma}^2_{\varepsilon}} = \frac{8.469}{8.469 + 149.76} = .054.$$

That is, after accounting for the first LSQ, about 5.5% of the variance in outcomes is due to the therapists at the average level of initial severity. Therapist variability was due almost entirely to intercept differences—that is, therapists differ on the mean outcomes that they obtain (again, correcting for first LSQs).

To interpret the results of the therapist variability, we used the following strategy. Because the difference between the regression coefficients calculated within therapist and between therapist was small, the error created by using residualized gain scores is negligible.[2] Most clinical trial analyses use such scores, which may or may not be justified in those studies, depending on the relationships found in the data. So, at this point, we repeated the analysis using the residualized gain scores and compared it with clinical trials to demonstrate the effects of this variability. Using the residualized gain scores calculated by regressing the last LSQ onto the first LSQ, we found

$$\hat{\sigma}^2_{\text{ther}} = \text{Therapist Variance Estimate} = 8.469$$

and

$$\hat{\sigma}^2_{\varepsilon} = \text{Patient Variance Estimate} = 154.35.$$

And thus the proportion of variability in outcomes due to therapists is given by

$$\frac{\hat{\sigma}^2_{\text{ther}}}{\hat{\sigma}^2_{\text{ther}} + \hat{\sigma}^2_{\varepsilon}} = \frac{8.469}{8.469 + 154.35} = .052,$$

which clearly is not very different than the more complex model allowing regressions to vary within and among therapists. We take 5% to be a good estimate of the variability in outcomes due to therapists, taking into account the initial score, in a managed care context.

[1] The equations for and a more extensive explanation of the various models are available from Bruce E. Wampold.

[2] The difference between within and between therapist slopes was determined by comparing models defined by the parameterization of the first LSQ (see Snijders & Bosker, 1999, Section 4.5). The parameter for the first LSQ, centered around the therapist mean first LSQ, was small (viz., .086), indicating that the within and between therapist slopes were comparable.

Variability in Outcomes Attributable to Therapists Illustrated

One way to understand therapist variance is to examine the consistency of therapist outcomes over time—in a sense, a cross-validation. To this end, we examined each therapist's caseload and divided it into two subsamples by time. For example, if a therapist had 10 patients in the sample, the first 5 seen were allocated to the first subsample, labeled *the predictor sample*, and the second 5 to the second subsample, labeled *the criterion sample*. The therapists were then placed into quartiles on the basis of their outcomes in the predictor subsample, using residualized gain scores as the criterion (negative residualized gain scores indicate better than predicted outcomes and vice versa). We then examined the outcomes of these top and bottom quartiles of therapists in the criterion subsample to determine the outcomes of the therapists identified as "best" and "worst" on the basis of previous performance. Of course, the more patients seen, the better the estimates of therapist outcomes. In Table 3, the results of this analysis are shown for therapists who saw at least 6 patients (3 in the predictor and 3 in the criterion subsamples) and 18 patients (9 in the predictor and 9 in the criterion subsamples) on the basis of the residualized scores.

The table also presents the proportion of patients who reliably changed (i.e., improved more than two standard errors of the differences from initial LSQ to final LSQ; Jacobson & Truax, 1991) and effect size for the patients (pretest minus posttest, divided by the standard deviation of the LSQ obtained in the norming study). Clearly those identified as effective therapists (the top quartile in the predictor sample) had better outcomes with their successive patients. The patients of the "best" therapists, in the cross-validated sample, had negative residualized gain scores (i.e., patients did better than expected), had a higher probability of displaying a reliable change than did the "worst" therapists (7%–13% greater, depending on the sample used), and produced pre–post effect sizes approximately twice as large as did the "worst" therapists. Thus, therapists identified by their performance in one time period continued to produce consistent results in subsequent time periods, demonstrating a stability of outcomes.

Therapist, Patient, and Treatment Variables

Four therapist variables in the data set might account for the amount of variability in outcomes attributable to therapists: degree, age, gender, and years of experience. Each therapist variable was added one at a time and together determine which, if any, of the therapist variables accounted for the variance among therapists over and above a model using an unconditional model (i.e., not conditional on therapist variables) based on residualized gain scores and fixing the slopes. The proportion of variance for the unconditional model was .051; when therapist variables were added the proportions of variance due to therapist ranged from .049 to .060, indicating that therapist variables added little to the understanding of therapist variance.

Table 3
Cross Validation of Therapist Effects (Mean Outcomes for the Criterion Sample)

Variable	Quartile 1 (best)	Quartile 4 (worst)
3 cases (483 therapists)		
Residualized change	−1.30	1.90
Proportion reliably changed	0.32	0.25
Effect size	0.43	0.23
9 cases (73 therapists)		
Residualized change	−1.81	2.30
Proportion reliably changed	0.35	0.22
Effect size	0.47	0.20

Note. Therapists were placed in quartiles on the basis of residualized gain scores of the predictor sample. The results in this table are means derived from the criterion sample.

The first patient variable examined was diagnoses. Because the distribution of diagnoses among therapists varied dramatically, modeling diagnosis within a multilevel framework proved difficult (i.e., solutions would not converge and estimates were unstable), we chose to ignore the therapists initially. The first analysis for diagnoses was a simple ANOVA, in which the diagnostic group was the independent variable and residualized scores were the dependent variable. To reduce the disparity in sample sizes, we included only diagnostic groups with more than 100 patients. The means and standard deviations are presented in Table 1. The omnibus F test was significant, $F(5, 5999) = 2.86$, $p = .014$. Adjustment disorder and anxiety disorders showed the most change, whereas bipolar disorder showed the least. Although the diagnostic group produced a statistically significant result, the size of the effect was small. In the present context, the effect size can be gauged by $\hat{\omega}^2$, as discussed by Hays (1994), which is an estimate of the population value of the proportion of variance explained by a fixed independent variable. For these data, $\hat{\omega}^2 = .0017$, indicating that less than 1% of the variance in outcomes was due to the diagnostic group to which the patient belonged (accounting for the initial degree of severity).

Although the differences in outcomes among the various diagnostic groups were small, it was possible that those therapists who had better outcomes were seeing patients with diagnoses that were most amenable to treatment. To rule out that possibility, we used the following strategy. First, we computed residualized gain scores within each diagnostic group (i.e., using a regression equation for patients within group), disregarding the therapist. Then, using these residualized scores, we calculated the estimate of the variance attributable to the therapist, as described previously (i.e., treating the therapist as a random factor). The following estimate of the proportion of variance attributable to the therapist was found:

$$\hat{\rho}_I = \frac{\hat{\sigma}^2_{\text{ther}}}{\hat{\sigma}^2_{\text{ther}} + \hat{\sigma}^2_{\varepsilon}} = \frac{8.60}{8.60+151.13} = .054.$$

This estimate approximates closely those of therapist effects obtained early, strongly suggesting that the diagnosis of the patient did not account for the differences among therapists observed in these data.

Mixed models were run with patient age and patient gender, the other two patient demographic variables available in the data set. These variables were considered as both patient and therapist variables, the latter by taking into account therapist differences in the age and gender of the patients they treated. Finally, the interactions of therapist and patient variables were considered. None of these analyses changed the proportion of variability in outcomes due to the therapist by more than .001.

We now turn to an interesting question about a patient variable that must be answered tentatively, given the nature of the data set: How does the administration of a psychotropic drug during the course of psychotherapy affect the variability in outcomes among therapists? On the basis of (a) the 1,083 patients who did not receive any medication during their episode of psychotherapy care and (b) the 586 who received medication concurrent with psychotherapy and used residualized gain scores, those on concurrent medications showed more benefit than those who did not receive medication ($M = -1.96$, $SD = 14.93$, for the medication condition; $M = 0.61$, $SD = 11.87$, for the nonmedicated condition), $t(1{,}667) = 3.84$, $p < .0001$. To understand this difference in the context of the individual therapists, we needed to restrict the sample to those who treated sufficient number of patients with and without medication to conduct analyses at the therapist level. Thus we analyzed data from 15 therapists who had at least three patients with concurrent medications and three patients with no medication (167 patients; only 1 medical doctor therapist provided both medications and psychotherapy).

Figure 2 shows the mean residuals for the 15 therapists disaggregated by whether their patients were receiving concurrent medication or no medication. Because negative residuals indicate better outcomes, the graph suggests that for the more effective therapists, their patients on medication did considerably better than did their patients not on medication, whereas for the therapists with poorer outcomes, patient outcomes appeared about equal. Further analyses of these data revealed the nature of this effect.

First, we examined the variance attributable to therapists for their patients not on medication and for those patients on medication, using the methods described above. For the patients not on medication, the proportion of variance due to therapist is given by the following:

$$\hat{\rho}_I = \frac{\hat{\sigma}^2_{\text{ther}}}{\hat{\sigma}^2_{\text{ther}} + \hat{\sigma}^2_{\varepsilon}} = \frac{7.82}{7.82+130.15} = .057,$$

This result is in the neighborhood of what has been obtained for therapists throughout. For patients on concurrent medication, the estimate of the proportion of variance attributable to therapist is given by the following:

$$\hat{\rho}_I = \frac{\hat{\sigma}^2_{\text{ther}}}{\hat{\sigma}^2_{\text{ther}} + \hat{\sigma}^2_{\varepsilon}} = \frac{92.84}{92.84+170.46} = .353.$$

Clearly, the proportion of variance in outcomes attributable to therapists treating patients on medication is greater than the proportion of variance in outcomes attributable to therapists who are treating patients not on medication. For the medication condition, 35% of the variance in outcomes was due to the therapist. This result is counterintuitive because the effects of medication should be independent of the administrator of the psychotherapy if the major benefits are due to the specific compounds rather than to the manner in which medication is given or to the nature of the concurrent psychotherapy.

For these 15 therapists, the appropriate mixed-model crossed design yielded the following effects: therapist main effect, $F(14, 137) = 4.41$, $p < .0001$; medication main effect $F(1, 14) = 3.23$, $p = .094$; and therapist–medication interaction: $F(14, 137) = 1.54$, $p = .105$. Because of the relatively low power and given the

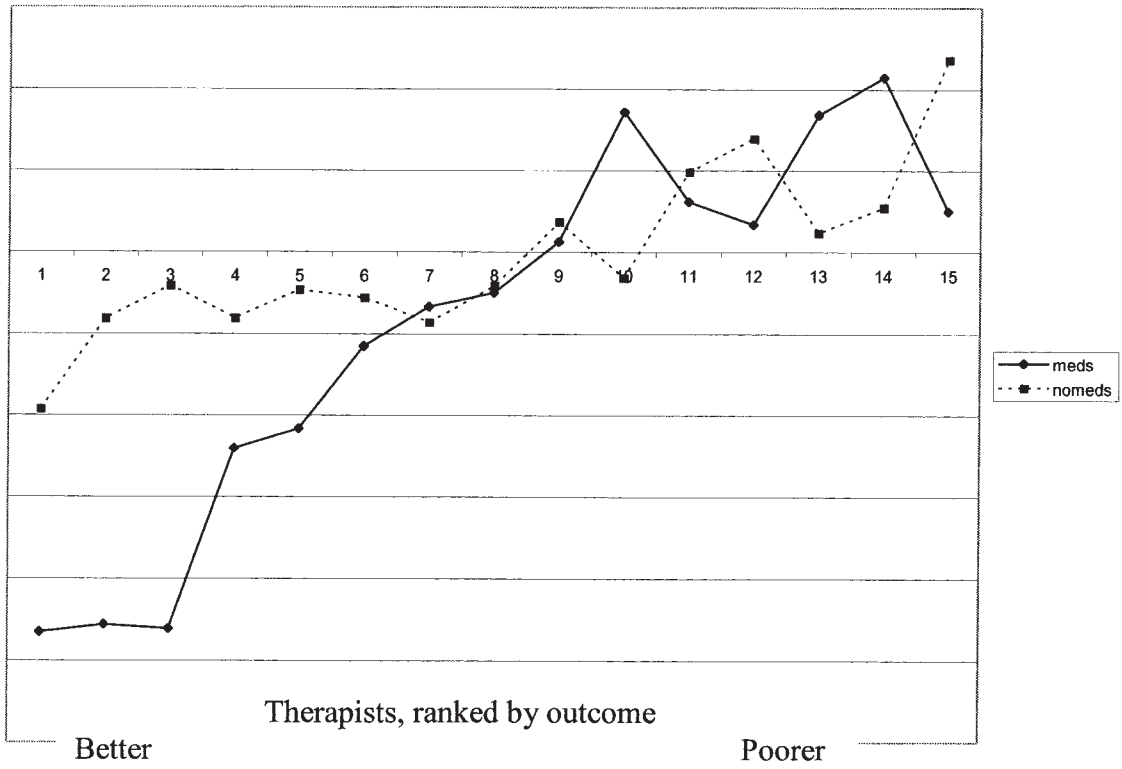

Figure 2. Outcomes (residualized gain scores) of 15 therapists for patients with concurrent medication (meds) and no medication (nomeds).

limited number of therapists and patients per therapist, neither the main effect for medication nor the interaction reached significance. In this reduced sample, the mean residuals were −4.02 for the concurrent medication condition (*SD* = 15.85) and −0.76 (*SD* = 11.73) for the medication condition. Again, it appears that the better therapist had considerably more success with medications relative to the poorer therapists, but this interaction effect did not reach significance either. However, these data suggest that the medication does not work independently of the therapist; that is, the medications did not produce a constant benefit that was independent of the psychotherapist. Caution needs to be exercised with regard to conclusions concerning medication effects because of the small sample sizes, and the results should be considered exploratory.

Discussion

Conclusions drawn from clinical trials of psychotherapy treatment are limited in the sense that their generalizability to practice settings is tenuous, given the many aspects of clinical trials that render the trial context different from the practice context (Westen et al., 2004). The primary purpose of the present study was to estimate the proportion of variability in outcomes attributable to the therapist in a managed care setting to compare this with the modal value found in clinical trials. We found that, taking into consideration the initial severity of the patient, about 5% of the variance in outcomes was due to the therapist, an estimate marginally lower than the 8% figure found in clinical trials. In the context of managed care, therapist age, gender, degree, and years of experience did not explain the variability among therapists, nor did the patient's age, gender, or diagnosis. Patients who received medication concurrent with psychotherapy showed greater change, although the amount of variance due to medication status was small in comparison with the variance that can be ascribed to therapists. Although the sample size was small and therefore the conclusion tentative, our results suggest the following: It appears that patients who take medication and see therapists who produce the best outcomes for patients without medication benefit more from the medication than do those patients who see therapists whose psychotherapy-only outcomes are poorer.

Compared with the 8% of variability in outcomes attributable to therapists that has been found in clinical trials, the 5% obtained here appears modest. Indeed, at first glance, one would be curious about why therapists in practice, who treat patients with various diagnoses, a wide range of severity, and significant comorbidity, using an unrestricted range of therapeutic approaches would produce less variability in outcomes than would therapists in clinical trials, in which, as it has been mentioned, the conditions are standardized. Perhaps the answer can be found in the nature of the patients. Recall that the intraclass correlation coefficient is the ratio of the variance attributable to therapists over the total variance in outcomes (i.e., the sum of the patient variance within therapists and the therapist variance). In clinical trials, the range of severity is restricted and the patients are homogenous in that their characteristics are constrained by the inclusion–exclusion criteria (e.g., all have the same diagnosis, are not suicidal, have limited comorbidity, are not taking psychotropic medications). Consequently, patient (error) variability in clinical trials is reduced, thus reducing the denominator of the intraclass correlation coefficient and increasing the proportion of variability in outcomes due to therapists. It is important to remember that the 5% detected in this study is a population estimate. If all therapists were equally effective, there would be observed variation among them because of sampling error; the 5% takes into account sampling error and is thus the variability in therapists over and above what would be expected by such sampling error. The variability among therapists in this sample translates into clinically meaningful differences in outcomes of patients in Year 2 on the basis of performance in Year 1 (see Table 3). It is worth noting that this finding may be the first in the literature in which therapists, who have been identified as being competent in one time period, empirically demonstrate superior outcomes in a future time period.

With the exception of medication, the therapist and patient variables did not account for much if any of the therapist variability. However, many patient and therapist characteristics were not measured in this study, and it could be that variability in outcomes among therapists was due to a biased distribution of these characteristics among therapists—for example, some therapists may see a greater proportion of patients who have poor prognoses. As well, important therapist or process variables, such as empathy or working alliance, were not assessed. Thus we were not able to identify what the better therapists possess or do that leads to their consistently better outcomes.

The results with regard to medication status are interesting and provocative. There is evidence that psychotherapy added to pharmacotherapy for patients with severe or persistent disorders increases the therapeutic benefits (Thase & Jindal, 2004).[3] The results of the present study, although limited by the small sample size of this particular analysis, suggest that the effects of combined treatments are dependent on the therapist. It appears that the patients of generally effective therapists benefit from the psychopharmacological treatments, whereas the patients of less effective therapists seem to benefit little, if any, from the medications. It may well be that effective therapists construct an expectation that the medications will be effective, increasing a placebo effect (Kirsch, 2005). In any event, the advice of Thase and Jindal (2004), with regard to combining psychotherapy and psychopharmacology, should be reiterated, given these findings: "Treatments that convey benefit only when provided by hand-picked, highly skilled therapists offer little value to patients treated in busy urban clinics or community mental health centers" (p. 760).

The importance of variability in outcomes attributable to therapists in practice settings raises many issues for the practice of psychology and the management of mental health services. A primary question is whether therapists who are consistently producing below average outcomes are aware of the fact that their patients are not progressing as well as expected. Psychotherapists appear to be subject to the same errors in judgment that are ubiquitous in most contexts (Baron, 2000; Turk & Salovey, 1988) and thus it would not be surprising to find that therapists are not particularly adept at identifying treatment success and failure. Indeed, Hannan et al. (2005) asked 48 therapists to assess treat-

[3] The mean initial severity of the medicated patients was greater than that of the nonmedicated patients, suggesting that combined treatments were appropriately being administered to those patients with more severe dysfunction.

ment progress and found that therapists do not recognize treatment deterioration. Exacerbating this problem is that therapists typically are not cognizant of the trajectory of change of patients seen by therapists in general. That is to say, they have no way of comparing their treatment outcomes with those obtained by other therapists.

Clinically, it would seem prudent for therapists to assess treatment outcomes and to have access to normative data. Providing feedback to therapists with regard to their outcomes vis-à-vis the average trajectory of change for patients with the same initial level of severity appears to increase the likelihood of positive outcomes (e.g., Lambert, Hansen, & Finch, 2001). However, it is rare for therapists in independent practice to collect outcome data or to have access to data for other therapists so that they can assess the effectiveness of their services. Similarly, organizations that manage care rarely collect or utilize outcome data. PBH, in an attempt to document the effectiveness of the services of their providers and to use their resources to increase benefits to patients, routinely collects outcome data and provides feedback to clinicians via letters generated in an automated manner by the clinical information system (Brown & Jones, 2005; Matumoto et al., 2003). Letters are generated if the patient's trajectory of change differs significantly from the expected course of recovery. These letters indicate that the probability of improvement remains high if the patient remains engaged in treatment. The therapist is encouraged to proactively address the risk of premature termination. There is a growing literature on outcomes-informed practice as a means to improve outcomes in practice (see Miller, Duncan, & Hubble, 2005).

Collection of outcome data and feedback relative to expected trajectory of change has the potential to provide information necessary to improve the outcomes of therapists with consistently below average results. However, if this feedback is insufficient to reduce the variability in outcomes among therapists, leaving some therapists who produce consistently below average results, what action should be taken? Several options are available, but controversial: Licensing boards could monitor outcomes and sanction poor performing therapists, payments to therapists could be contingent on outcomes, or managed care could steer referrals to better performing therapists.

A number of limitations are inherent in the research reported here. First, because the data were collected in a naturalistic context, experimental manipulations were precluded. Most detrimental to the present study is that patients were not randomly assigned to therapists; although initial severity and several patient and therapist variables were examined, the influence of biased assignment because of unmeasured variables cannot be identified nor corrected. Also, the naturalistic nature of the data necessitated a reasonable, although arbitrary, definition of an episode of care. Further, despite treating therapists as a random factor, the results are generalizable only to therapists similar to those on the provider panel of PBH with patients similar to those insured by PBH (see Serlin et al., 2003). Moreover, the type of treatment provided by the PBH therapists was unknown so it is not possible to determine whether the variability in outcomes is due to characteristics of the therapist or the effectiveness of the treatments delivered. Although the naturalistic nature of the present study extends knowledge of therapist variability from the clinical trial context, the results are limited to a managed care context where covered individuals have relatively generous mental health benefits and are employed or belong to a family in which a member is employed. Generalizability is further restricted in that not all patients participated, an issue particularly apparent in the first year of this study. Finally, the outcome measure assessed global functioning, and it is unclear whether variability in outcomes attributable to therapists would differ had measures specific to symptomology associated with each patient's disorder been used, although it should be noted that therapist variability in clinical trials does not seem to be related to the specificity of the measure (Kim et al., in press).

Future research on variability in outcomes attributable to therapists is needed to address two fundamental questions: What are the therapist characteristics and actions (including the treatment delivered) that account for variability among therapists? How can the benefits provided to patients by the therapists who achieve less than expected outcomes be improved?

References

Baron, J. (2000). *Thinking and deciding* (3rd ed.). Cambridge, England: Cambridge University Press.

Beck, A. T., Ward, C., Mendelson, M., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 6,* 561–571.

Beutler, L. E., Malik, M., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, S., & Wong, E. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 227–306). New York: Wiley.

Blatt, S. J., Sanislow, C. A., Zuroff, D. C., & Pilkonis, P. A. (1996). Characteristics of effective therapists: Further analyses of data from the National Institute of Mental Health treatment of depression collaborative research program. *Journal of Consulting and Clinical Psychology, 64,* 1276–1284.

Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vacarro, J. (2001). Pushing the quality envelope: A new outcomes management system. *Psychiatric Services, 52,* 925–934.

Brown, G. S., & Jones, E. R. (2005). Implementation of a feedback system in a managed care environment: What are patients teaching us? *Journal of Clinical Psychology/In Session, 61,* 99–110.

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

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

Derogatis, L. R. (1977). *The SCL-90 manual: Scoring, administration, and procedures for the SCL-90.* Baltimore: Johns Hopkins School of Medicine, Clinical Psychometrics Unit.

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

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology/In Session, 61,* 1–9.

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Holt, Rinehart, & Winston.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasensor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56,* 885–892.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2001). Therapists, therapist variables, and cognitive–

behavioral therapy outcomes in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69,* 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. (in press). Therapist effects in psychotherapy: A random effects modeling of the NIMH TDCRP data. *Psychotherapy Research.*

Kirsch, I. (2005). Placebo psychotherapy: Synonym or oxymoron. *Journal of Clinical Psychology, 61,* 791–803.

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Murish (Ed.), *Use of psychological testing for treatment planning and outcome assessment* (3rd ed., Vol. 3, pp. 191–234). Mahway, NJ: Erlbaum.

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., Hatfield, D. R., Vermeersch, D. A., Burlingame, G. M., Reisinger, C. W., & Brown, G. S. (2001). *Administration and scoring manual for the LSQ.* Salt Lake City, UT: American Professional Credentialing Services and Van Nuys, CA: PacifiCare Behavioral Health.

Matumoto, K., Jones, E., & Brown, J. (2003). Using clinical informatics to improve outcomes: A new approach to managing behavioral healthcare. *Journal of Information Technology in Health Care, 1,* 135–150.

Miller, S. D., Duncan, B. L., & Hubble, M. A. (2005). Outcome-informed clinical work. In J. C. Norcross & M. R. Goldfried (Eds.), *Handbook of psychotherapy integration* (2nd ed., pp. 84–102). New York: Oxford University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

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 Joorman (2003). *Psychological Methods, 8,* 524–534.

Siemer, M., & Joormann, J. (2003). Power and measures of effect size in analysis variance with fixed versus random nested factors. *Psychological Methods, 8,* 497–517.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Thase, M. E., & Jindal, R. D. (2004). Combining psychotherapy and psychopharmacology for treatment of mental disorders. In M. L. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 743–766). New York: Wiley.

Turk, D. C., & Salovey, P. (Eds.). (1988). *Reason, inference, and judgment in clinical psychology.* New York: Free Press.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74,* 242–261.

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings.* Mahwah, NJ: Erlbaum.

Wampold, B. E., & Serlin, R. C. (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425–433.

Weissman, M. M., & Bothwell, S. (1976). Assessment of patient social adjustment by patient self-report. *Archives of General Psychiatry, 33,* 1111–1115.

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130,* 631–663.

Received July 25, 2004
Revision received March 16, 2005
Accepted March 22, 2005 ■