An Analysis of Therapist Treatment Effects: Toward Providing Feedback to Individual Therapists on Their Clients' Psychotherapy Outcome

▼

John C. Okiishi, Michael J. Lambert, Dennis Eggett, Lars Nielsen, and David D. Dayton
Brigham Young University

▼

David A. Vermeersch
Loma Linda University

This study examined data collected on over 5,000 clients seen by 71 therapists over a 6-year period in a University Counseling Center. Clients were given the Outcome Questionnaire-45 (OQ-45) on a session-by-session basis to track their treatment response. Data were also collected on therapists' theoretical orientation, years of experience, gender, and type of training. Data were analyzed using hierarchical linear modeling (HLM) to see if general therapist traits (i.e., theoretical orientation, type of training) accounted for differences in clients' rate of improvement. Data were then analyzed, again using HLM with therapists as a fixed effect, to compare individual therapists to see if there were significant differences in the efficiency of treatment. In addition, pre- minus posttest OQ-45 scores were examined to see if there were differences in the overall outcome of clients. There was a significant amount of variation among therapists' clients' rates of improvement. Therapist feedback reports were generated to summarize client outcome for individual providers in contrast to center averages and in an attempt to improve client outcome. © 2006 Wiley Periodicals, Inc. J Clin Psychol 62: 1157–1172, 2006.

Keywords: Outcome Questionnaire-45; therapist effects; psychotherapy outcome; therapist effectiveness; outcomes management

Correspondence concerning this article should be addressed to: Michael J. Lambert, Department of Psychology, 272 TLRB, Brigham Young University, Provo, UT 84602; e-mail: michael_lambert@byu.edu

JOURNAL OF CLINICAL PSYCHOLOGY, Vol. 62(9), 1157–1172 (2006)

Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jclp.20272

Outcome research in the last decade has extensively focused on the effects of specific treatments for specific disorders, so-called clinical trials. Researchers employing this methodology typically eliminate the individual therapist as an important factor in client outcome in their primary analysis of treatment effects by expending considerable resources to diminish variability in outcomes that could be attributed to the therapist through selection, training, supervision, and the use of treatment manuals (Lambert & Ogles, 2004; Wampold, 2001). This increases the likelihood of finding effects due to treatments and independent of the therapists who offer them. Such design tactics make perfect sense if the goal of a study is to detect and maximize the contribution of theory-based interventions on client outcome. Such studies are used as a basis for recommendations (evidence-based clinical practice guidelines) that are intended to eventually improve the quality of clinical services.

Despite attempts to eliminate the therapist as an important outcome variable, some evidence suggests that the individual therapist can have a substantial impact on client outcome. For example, Luborsky, McClellan, Woody, O'Brien, and Auerbach (1985) studied outcome for opiate addiction in a clinical trial comparing cognitive behavior therapy, supportive-expressive psychotherapy, and drug counseling. The outcome effect size varied across the nine therapists, ranging from .13 to .79 despite the usual attempts to minimize variance across therapists. They concluded that "profound differences were discovered in the therapists' success with the clients in their case load" (p. 602).

Along similar lines, data from other clinical trials have been reanalyzed with the intent of quantifying the contribution of the individual therapist to outcome. Among these, the NIMH Collaborative Depression Study has been examined, and depending on type of analysis, no differences have been found between types of psychotherapy in the face of moderate effects attributable to specific therapists (Elkin, Falconnier, Martinovitch, & Mahoney, 2006). Meta-analytic reviews have documented similar, though less dramatic, findings across a variety of treatments and methodologies. For example, Crits-Christoph and Mintz (1991) meta-analyzed data from 10 clinical trials and noted the surprisingly high contribution of therapists to outcome, suggesting that the size of their contribution posed a challenge to attributing differential treatment effects to the specific therapies that were under study and the need to take these effects into account when analyzing outcomes. Therefore, even in the context of clinical trials, where considerable effort is aimed at diminishing the effects that can be attributable to individual therapists, such effects are still often substantial. At the same time, Crits-Christoph and Gallop (2006) have cautioned about overstating the effects of individual therapists in clinical trials as no effects attributed to therapists have been reported in some studies.

Therapist effects have also been reported in the context of effectiveness research conducted within routine care. Such effects might be expected to be larger than those found in clinical trials because practice-based research typically uses available therapists without extensive use of selection, training, and rigorously controlled treatments. Crits-Christoph and Gallop (2006) estimated that such studies appear to produce small-to-moderate therapist effects. Within this practice-based context, Okiishi, Lambert, Nielsen, and Ogles (2003) analyzed the psychotherapy outcome of 1,841 clients seen by 91 therapists. The outcome by therapist showed considerable variability, with the most-effective therapist's clients showing both rapid and substantial treatment response, whereas the least-effective therapist's clients showed an average worsening in functioning. In contrast, there was no evidence that the same degree of variability and superiority could be found for outcomes by treatment orientation.

A major disadvantage for drawing conclusions about the relative contribution of variables to client outcome in research conducted outside of clinical trials is that many of

the variables of interest cannot be adequately operationalized. For example, the diagnosis given to a client is made by the treating clinician without the reliability and validity checks that are used in clinical trials, and treatments are not provided with any fidelity checks or session recordings. Because of these limitations, the failure to find a relationship between the type of therapy and outcome in routine practice research is not as meaningful as it would be in clinical trials. In the context of this practice-based research, more pragmatic questions take precedence.

Whether making a referral, hiring a new therapist at a group practice, selecting a therapist for inclusion in a panel of providers, or deciding who to go to for one's own therapy, judgments about the best providers are made on a daily basis. These judgments typically occur in the absence of empirical information about clients' treatment response to individual therapists (Saptya, Riemer, & Bickman, 2005). Little effort has been expended on using actual treatment outcomes based on the effects of the individual therapist to directly improve client outcome although such data could dramatically influence the delivery of effective care (Orlinsky & Howard, 1980; Strupp, 1980). Hesitancy to use such information is hardly limited to behavioral health care. Millenson (1997) has documented a plethora of examples in the field of medicine where resistance to analyzing outcomes within hospitals and physicians has had disastrous consequences. This history includes the publication of important findings on the variability of outcomes across hospitals and physicians and the regular failure of institutions and the researchers involved to make use of such information to directly improve patient outcomes. Surprisingly, physicians are often disinterested in or even intransigent to feedback.

Variability attributable to individual therapists in routine care provides an opportunity to directly manage and improve outcomes in specific clinics. Examples of using outcome data for this purpose are beginning to emerge and are proffered as a quality management tool. For example, Brown, Lambert, Jones, and Minami (2004) ranked 2,459 individual clinicians based on their clients' outcomes during a 2-year period. Clinicians were included in the analysis if they had at least 10 cases with change scores in the baseline period. The clinicians were sorted after adjusting for case mix based on their clients mean intake scores and then assigned to quartiles. Clinicians in the top quartile and bottom quartile were compared. At-risk clients seen by the top-quartile clinicians averaged greater change while tending to utilize fewer sessions. In contrast, the at-risk cases treated by bottom-quartile clinicians continued to show deterioration despite averaging more sessions at the final assessment. PacifiCare Behavioral Health (Laguna Hills, CA) has used such data to reward superior outcomes on an annual basis by publicly recognizing the most effective clinicians (or group practices) and offering each a cash award of $1,000.

In the studies by Brown et al. (2004) and Okiishi et al. (2003) the number of clients seen by each therapist was relatively small (as few as 14 per therapist) raising questions about their reliability and usefulness for practical decision making. The current study was undertaken for similar purposes and was aimed at improving the quality of care at a university-based clinic but with larger sample sizes within therapists and by taking into account several important variables. Past research has clearly demonstrated the most important predictor of final outcome is initial level of client severity (Garfield, 1994). Prior investigation within our clinic indicated a relationship between severity as measured by psychological testing and diagnosis as well as diagnostic comorbidity, with higher scores associated with more serious diagnostic classifications (e.g., mood disorder vs. adjustment) and multiple diagnoses (Lueck, 2004). Once severity, based on test score, was accounted for, other client variables, such as diagnosis added little to predicting outcome. Thus, in the present study initial level of disturbance was examined across therapist caseload.

Second, characteristics of the treatment (theoretical orientation) and therapists (gender, experience, professional background) that might be related to treatment success were examined. Once these variables were analyzed, this study examined differences in the rate of client-treatment response (efficiency), the size of the treatment response (effectiveness), differences between therapists in terms of the number of sessions they provided clients, and differences in the degree to which clients within therapist caseloads met criteria for clinically significant change. The current study was undertaken with the intent of using the information obtained to improve care within the context of a single clinic, and in the spirit of helping clinicians to become more knowledgeable about the variability in client outcome that could be attributed to individual therapists.

Method

Participants

Clients. The client sample for this study consisted of college students seen at a large university counseling service for individual psychotherapy over a 6-year period. Treatment was available to fulltime students of the university. Clients at the center presented with a wide range of problems from simple homesickness to personality disorders. Clients were initially seen for a brief intake interview and then assigned to a particular therapist based on therapist availability and client needs. No experimental control was exercised over this routine procedure. There were no session limits imposed. The range of sessions in this initial sample was 1 to 203, with a mean of 8.74 ($SD = 14.3$).

Although 11,736 clients were seen at the center over the 6-year period of data collection, the statistical method being used for this study required at least three data points (a pretest and two additional measurements), so individuals with less than three treatment sessions, including the intake, were not included in the sample. This selection criterion yielded a data set of 7628 clients, who had been seen for 64,103 sessions in total. The most common diagnoses in the final data set were mood disorders ($n = 1{,}961$, 36.14%), anxiety disorders ($n = 1{,}200$, 22.11%), and adjustment disorders ($n = 913$, 16.83%).

Therapists. One hundred forty-nine therapists contributed data to the entire data pool of 7,628 clients. In our previous study (Okiishi et al., 2003), we set a minimum number of clients per therapist at 15. To gain a more accurate assessment of client improvement, we set the minimum number of clients at 30 for this study. Using this criterion, as well as the three data-point minimum described above, 71 therapists who had seen 6,499 clients were left in the sample. The therapists in this final data set had seen an average of 92 clients, for an average of 9.71 sessions. Data were also collected on a variety of therapist variables: level of training (preinternship, internship, and postinternship), type of training (clinical psychology, counseling psychology, social work, marriage and family therapy), sex (male, female), and primary theoretical orientation (cognitive–behavioral, behavioral, humanistic, psychodynamic). The modal therapist was a male, licensed, counseling psychologist with a doctorate, who identified his primary theoretical orientation as cognitive–behavioral.

Measure

Client progress in this study was tracked using the Outcome Questionnaire (OQ-45), a 45-item self-report measure developed specifically for the purpose of tracking and assessing client outcomes in a therapeutic setting. The OQ-45 is a well-established instrument

that has been validated across the country and across a broad range of normal and client populations. Lambert et al. (2004) reported an internal consistency for the OQ-45 of .93 and a 3-week test–retest value of .84 both of which are considered adequate. Concurrent validity figures were calculated by comparing the OQ-45 total score with total scores from other measures including the Symptom Checklist-90 (SCL-90; Derogatis, 1977), Beck Depression Inventory (BDI; Beck, Steer, & Garbin, 1988), Zung Depression Scale (Zung, 1965), and the State-Trait Anxiety Inventory (STAI; Spielberger, 1983). All of the concurrent validity figures with the OQ-45 and these instruments were significant at the .01 level with a range of *r*'s from .50 to .85. Most important, the OQ-45 has been shown to be sensitive to the effects of interventions on client functioning (Vermeersch, Lambert, & Burlingame, 2000; Vermeersch et al., 2004).

The OQ-45 is scored using a 5-point scale (0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *frequently*, 4 = *almost always*), which yields a possible range of scores from 0 to 180. High scores on the OQ-45 indicate more distress and as clients improve scores decrease. Although not used in this study, the OQ-45 has three subscales that measure quality of interpersonal relations, social role functioning, and symptom distress. The total score, which provides a global assessment of functioning, was used in this study.

Using formulas developed by Jacobson and Truax (1991), clinical and normative data for the OQ-45 were analyzed by Lambert et al. (2004) to provide cutoff scores for the Reliable Change Index (RCI) and clinically significant change. Clients who change in a positive or negative direction by at least 14 points are regarded as having made "reliable change." This degree of change exceeds measurement error based on the reliability of the OQ-45 and is one of two criteria posited by Jacobson and Truax (1991) as indicative of clinically meaningful change. The second criterion requires movement from a score typical of a dysfunctional population to a score typical of a functional population (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). The cutoff on the OQ-45 for marking the point at which a person's score is more likely to come from the dysfunctional population than a functional population has been estimated to be 64. When a client's score falls at, or below, 63 it is concluded that this client's functioning is similar to a nonclient's level of functioning at that point in time. Passing this cutoff (from dysfunctional to functional) is the second criterion posited by Jacobson and Truax (1991) as an indicator of clinically significant change.

Clients who show reliable change and pass the cutoff are considered "recovered"; those who only show reliable change are considered "improved." Clients who do not change more than 14 points in a positive or negative direction are considered "no change," and clients who worsen by 14 points are considered "deteriorated." Support for the validity of the OQ-45's reliable change and clinical significance cutoff scores have been reported by Lunnen and Ogles (1998), Beckstead et al. (2003), as well as Bauer, Lambert, and Nielsen (2004). This research suggests that the Jacobson–Truax formulas provide a sound basis for estimating cutoff scores, and that classification of change based on other measures and measurement sources results in considerable consensus on the individual case.

Procedures

In addition to providing information about the progress of the clients, each OQ-45 contained information making it possible to identify the client, the therapist, and the date of the session. As often as possible, the OQ-45 was administered to clients before each session but the statistical methods being used in this study did allow for missing values and collection of data at variable intervals. To protect therapist identities, a nontherapist

consultant randomly assigned each therapist a number so that identifying individual therapists would be impossible by viewing the data set. Client identities were removed prior to data analysis.

Data were analyzed using hierarchical linear modeling (HLM). The computer software used for this analysis was SAS for Windows (version 9.1; SAS Institute, Inc., Cary, NC). Hierarchical linear modeling has been demonstrated to have a number of advantages over other multivariate repeated measures methods (Bryk & Raudenbush, 1992; Singer, 1998). Hierarchical linear modeling is ideal for a naturalistic study such as this in that it accounts for missing and erratic data. As long as three or more data points are available, HLM allows for computation of estimates for missing data. Hierarchical linear modeling is a "nested regression" and provides a line of best fit with a slope and intercept.

An initial HLM was performed taking into account therapist variables (level of training, primary theoretical orientation, type of training, and gender). This was done to answer the question: "Do available therapist variables account for differences in clients' outcome?" Given past effectiveness research, it was not anticipated these variables would contribute significantly to the outcome of a client. However, in case that they did, this possibility needed to be considered before drawing conclusions about therapist outcomes. Furthermore, an ANOVA was performed on a client's initial OQ-45 scores by the therapist to answer the question: "Do some therapists see clients whose average initial disturbance is greater than other therapists?"

Following this initial check of therapist variables, we employed two ways of ranking therapists. Our first method was to use HLM slopes to examine the modeled rate at which clients' OQ-45 scores decreased over sessions of psychotherapy. This gave an indication of the average amount of change a therapist's clients improved per session of psychotherapy. Once these slopes had been computed, therapists were rank ordered by HLM slope with the steepest slope being "#1" and the shallowest slope being "#71." Because HLM slopes look at rapidity of symptom alleviation this method was considered best for determining the "efficiency" of a therapist.

As HLM slope gives important information about the speed at which therapists' clients improve (efficiency), it also provides a modeled picture of outcome in which data from any specific case is affected by data from all other cases within a therapist's caseload. In such models the growth curve of a therapist's cases shrink towards an average slope. Slopes are also plotted on a common, uniform, number of sessions that is somewhat arbitrarily chosen but that does not directly take into account the fact that some therapists have longer or shorter average number of sessions they provide to clients.

As another method of ranking therapist effects, change scores were computed for each client by subtracting the client's last OQ-45 score from their initial OQ-45 score. These change scores were averaged for all clients by therapist. This provided the actual amount of change experienced by clients seen by a therapist rather than a change index based on a statistical model of change (line of best fit as computed by the HLM analysis). Once these average scores had been computed, therapists were again ranked from most to least effective. The disadvantage of this index of change (unlike HLM) is that it is based on only two scores; however, the advantage is that it expressed the self-reported change of clients from the beginning to the end of therapy rather than an estimate of the slope of change.

In the final analysis therapist ranking on the two indexes of outcome were averaged and a final ranking of the 71 therapists was accomplished. The outcome of the top 10% ($n = 7$) of therapists and the bottom 10% ($n = 7$) was then examined to judge the clinical significance of the differences in client outcome in terms of Jacobson and Truax (1991) criteria: "recovered," "improved," "no change," or "deteriorated." A chi-square analysis

was performed to determine if there were statistically significant differences between top therapists' and bottom therapists' clients in categorized outcome. It was hoped that the selection of these extreme groups might lead to the examination of the practices of the most-effective psychotherapists.

Results

Therapist Variables

The initial HLM analysis of therapist data was aimed at seeing if therapists differed from one another on outcomes based on four therapist variables: type of training, amount of training, theoretical orientation, and gender. The results of this analysis indicated that the type of training (counseling psychology, clinical psychology, marriage and family therapy, social work), the years of training (preinternship, internship, postinternship), theoretical orientation (behavioral, cognitive–behavioral, humanistic, psychodynamic), and gender did not contribute significantly to the speed with which clients improved. Results are shown in Table 1. These findings suggest that any differences found between individual therapists' clients outcome trajectory were the result of variables not assessed in the current study, or through interactions between studied variables and other variables. In short, client recovery could not be shown to be a function of differences in therapist gender, professional training, professional experience, or theoretical orientation. This finding held up when pre- and postchanges were examined rather than the slope of change.

Initial Outcome Questionnaire-45 Scores

An ANOVA was performed on client's initial OQ-45 scores by therapist. Results of this procedure indicated that there were no significant differences between therapists' clients' initial OQ-45 scores ($F_{71,\ 6428} = 1.22$, $p > .05$). In other words, therapists did not have unequal caseloads based on initial OQ scores, suggesting that significant differences in outcome between therapists would not be due to variation in difficulty of caseload. The average initial OQ score for all clients was 66.61.

Differences Between Therapists

The HLM analysis indicated that clients' HLM lines had an average intercept (which can be interpreted as an estimate of pretreatment severity of distress) of OQ-45 = 63.26, with

Table 1
Hierarchical Linear Modeling (HLM) With Experience Level, Sex, Level of Training, and Theoretical Orientation as Predictors

Fixed effect	*df*	*F* Value	*P*
Level of training	2	0.30	0.74
Sex of therapist	1	2.78	0.10
Type of training	3	1.56	0.20
Primary theoretical orientation	3	1.62	0.18

a mean slope of −.75. The negative slope indicates an average decrease in OQ-45 points per session (i.e., a lessening of endorsed client distress) over the course of psychotherapy.

Hierarchical linear modeling slopes were also generated based on all of the clients in each therapist's caseload to compare therapists' outcomes to each other and to the general growth curve for the Center. Hierarchical linear modeling intercepts, slopes, and ranks for all the therapists in the sample are shown in Table 2. The therapists in Table 2 are ordered from 1 to 71 on the basis of their composite (slope and pre- and postchange) ranking. For purposes of reference, this composite ranking also serves as the therapist ID. The HLM analysis indicated that therapists' clients differed significantly on their rate of change. The therapists' growth curve slopes (i.e., rate at which clients' growth curves moved in a negative direction, indicating less-endorsed symptomatology) showed a wide range of variability ($F_{71,\ 3021} = 12.9$, $p < .001$). This finding suggests differential rates of change for clients depending on which therapist they saw. At the extremes, slopes by therapist ranged from an average of −2.07 OQ-45 points dropped per session (therapist #1, HLM rank = 1) to −.15 OQ-45 (therapist #66, HLM rank = 71) points dropped per session.

If all therapists saw their clients for approximately the same number of sessions the slope of change would provide a relatively accurate picture of client outcome and individual differences related to therapists. However, it is important to note that therapists also differed in the average number of sessions that they saw clients. As shown in Table 2, the extremes of the average number of sessions ranged from 22.03 for therapist #57, to 4.94 for therapist #5. Differences in the number of sessions therapists saw clients require that therapist effects be examined by analyzing pre- and postchange as well as the slope of change.

Pre- and Postchange Scores

The results of judging client outcome through the use of pre-and postchange scores are presented in Table 2. As can be seen, using pre- and postchange to examine therapist effectiveness, the average client seeing therapist #23 (whose 46 clients averaged the greatest amount of improvement), improved by 14.93 OQ-45 points over about 11 treatment sessions. In contrast, the average client seeing therapist #67 (whose 165 clients averaged the least amount of improvement), improved by only 2.66 points over an average of 12 sessions of psychotherapy. Similarly, the 2nd ranked therapist (#9) saw 39 clients an average of over 10 sessions who improved an average of 14.61 OQ-45 points, while the 70th ranked therapist (#68) saw 43 clients an average of 12 sessions who improved by only 4.23 OQ-45 points.

As can be seen in Table 2, when pre- to postchange was used as an index of outcome, therapist rank ordering varied from HLM ranking. For example, therapist #1, whose clients had the steepest HLM curve of all the therapists in the sample, dropped in ranking from 1st to 6th when pre- to posttest ordering was used. Conversely, therapist #23 went from being ranked 53rd using the HLM slope, to 1st when using pre- and postdifference scores. The correlation between slope of change and pre- and postchange was $r = .46$ ($p = .001$), and while statistically significant, suggests the two indices do not share a great deal of common variability. Given the goal of using the results of the current study as a basis for giving therapists feedback about the progress of their clients in relation to the progress at the center as a whole, and the fact that feedback on the two indexes of outcome did not provide the same rank ordering, it was decided that therapists should be given information about their rank on both efficiency and effectiveness, as well as a

Table 2
Composite Therapist Ranking, Slopes, Sessions, and Mean Pre- minus Post-OQ-45 Score Across 71 Therapists in a University Outpatient Clinic

Therapist composite rank & ID/*N*[a]	Average intake OQ-45 total	Average # of sessions	Pre-post-change/**rank**	HLM slope/**rank**
1/**43**	67.74	5.26	13.55/**6**	2.07/**1**
2/**39**	65.36	5.72	13.46/**7**	2.00/**2**
3/**86**	71.47	7.30	13.89/**5**	1.49/**6**
4/**84**	70.56	11.9	14.19/**4**	1.20/**8**
5/**34**	65.59	4.94	12.79/**10**	1.89/**3**
6/**36**	65.97	12.58	14.19/**3**	1.08/**13**
7/**85**	66.84	7.64	12.13/**14**	1.38/**7**
8/**36**	67.08	10.47	12.17/**13**	1.17/**9**
9/**39**	67.41	11.85	14.61/**2**	0.95/**24**
10/**84**	70.50	6.45	10.76/**23**	1.51/**4**
11/**34**	67.62	8.91	12.52/**12**	1.06/**15**
12/**36**	66.00	8.77	11.14/**19**	1.09/**12**
13/**75**	68.89	10.10	13.03/**9**	0.98/**23**
14/**171**	69.70	12.35	11.60/**16**	1.02/**21**
15/**49**	68.90	8.06	11.06/**21**	1.06/**17**
16/**174**	71.11	7.89	10.03/**29**	1.15/**10**
17/**35**	68.46	9.00	12.70/**11**	0.92/**28**
18/**62**	59.97	8.53	11.11/**20**	1.03/**20**
19/**53**	69.87	12.58	13.36/**8**	0.82/**34**
20/**55**	64.98	7.20	9.58/**32**	1.13/**11**
21/**227**	67.66	7.71	9.46/**34**	1.06/**16**
22/**97**	70.39	8.32	10.77/**22**	0.88/**30**
23/**46**	66.41	11.26	14.93/**1**	0.66/**53**
24/**378**	69.45	9.03	10.42/**25**	0.82/**35**
25/**167**	65.07	7.55	8.78/**42**	1.04/**19**
26/**119**	68.40	7.55	9.13/**40**	1.01/**22**
27/**31**	56.52	7.34	10.36/**26**	0.80/**36**
28/**67**	66.84	7.91	11.51/**17**	0.75/**46**
29/**41**	63.68	15.88	9.40/**37**	0.94/**26**
30/**51**	69.43	10.27	10.69/**24**	0.78/**40**
31/**42**	67.48	7.36	9.29/**39**	0.95/**25**
32/**41**	69.22	7.10	5.93/**63**	1.50/**5**
33/**38**	65.42	6.50	7.53/**54**	1.04/**18**
34/**138**	70.10	7.46	9.42/**36**	0.80/**38**
35/**48**	62.17	14.79	10.20/**27**	0.74/**47**
36/**53**	61.81	8.15	8.41/**48**	0.92/**27**
37/**32**	67.38	6.85	9.72/**31**	0.76/**44**
38/**196**	68.94	9.35	10.11/**28**	0.71/**50**
39/**50**	62.44	14.66	11.32/**18**	0.57/**61**
40/**30**	61.87	7.67	8.74/**43**	0.80/**37**
41/**115**	62.32	14.00	5.23/**66**	1.08/**14**
42/**51**	66.37	6.80	7.67/**52**	0.89/**29**
43/**56**	65.73	6.07	7.98/**51**	0.88/**31**
44/**130**	66.20	5.05	12.02/**15**	0.38/**68**
45/**274**	70.27	10.00	8.45/**46**	0.79/**39**
46/**171**	64.02	9.57	8.70/**44**	0.77/**42**
47/**327**	66.10	11.5	8.26/**50**	0.77/**41**
48/**159**	66.87	11.43	9.57/**33**	0.59/**58**
49/**154**	69.32	12.03	6.36/**61**	0.85/**32**
50/**46**	65.70	9.39	9.39/**38**	0.65/**55**

(*continued*)

Table 2
Continued

Therapist composite rank & ID/*N*[a]	Average intake OQ-45 total	Average # of sessions	Pre-post-change/**rank**	HLM slope/**rank**
51/**115**	64.79	12.51	8.33/**49**	0.75/**45**
52/**53**	70.26	8.50	9.87/**30**	0.54/**65**
53/**32**	65.16	13.22	5.72/**64**	0.83/**33**
54/**54**	63.35	8.44	6.82/**59**	0.77/**43**
55/**94**	67.62	11.97	9.44/**35**	0.42/**67**
56/**178**	69.29	11.26	8.93/**41**	0.55/**63**
57/**38**	66.68	22.03	7.58/**53**	0.67/**52**
58/**57**	67.56	8.09	6.54/**60**	0.73/**48**
59/**130**	68.84	8.80	6.97/**57**	0.69/**51**
60/**43**	68.53	9.63	8.63/**45**	0.49/**66**
61/**48**	64.83	9.81	6.85/**58**	0.65/**54**
62/**32**	62.15	9.25	8.44/**47**	0.34/**69**
63/**167**	68.13	17.63	4.72/**69**	0.72/**49**
64/**318**	63.64	8.45	5.98/**62**	0.57/**60**
65/**44**	63.82	8.82	7.28/**56**	0.27/**70**
66/**50**	62.56	8.22	7.36/**55**	0.15/**71**
67/**165**	67.08	12.15	2.66/**71**	0.62/**56**
68/**43**	66.46	12.19	4.23/**70**	0.61/**57**
69/**39**	66.97	9.18	4.87/**68**	0.59/**59**
70/**45**	68.89	13.13	5.67/**65**	0.55/**64**
71/**69**	65.39	10.41	5.23/**67**	0.56/**62**

[a]***N*** = Number of patients evaluated during study period.

composite ranking that combined the two pieces of information. The composite ranking is displayed as the therapist ID # in Table 2.

Average Ranking

Examining the data provided in Table 2 one can see that the typical client seeing the top-ranked therapist (#1, $n = 43$) could expect to begin treatment in the dysfunctional range (64 and above) on the OQ-45, be seen for about five sessions, drop 14 points on the OQ-45, and cross the clinical significance cutoff line of the OQ-45 (64/63). In short, clients seeing this therapist could, on average expect to meet clinical significance criteria for "recovery." At the other end of the therapeutic spectrum, a typical client seeing the worst-ranked therapist (#71, $n = 69$), would be expected to enter treatment slightly less disturbed, be seen for nearly twice the number of sessions (10.41), and improve by only 5 points on the OQ-45. On average, clients seen by this therapist would *not* be expected to leave therapy having experienced clinically significant change, let alone reliable improvement.

Using the average ranking as the estimate of therapist effectiveness, the top and bottom 10% of therapists were identified to more generally describe the range of client outcome at the extremes of the distribution. The seven most effective therapists saw their clients an average of 7.91 ($SD = 3.13$) sessions, had an average HLM slope of -1.59 ($SD = .40$) points per session and a pre- and postaverage change score of -13.46

(SD = .76). The seven least-effective therapists saw their clients an average of 10.59 (SD = 1.92) sessions, had an average HLM slope about 3 times less (−.48 points per session; SD = .19), and a pre-post-average change score of −5.33 (SD = 1.66), about one third that of the top group.

Outcomes of clients seen by these most- and least-effective therapists were classified according to the clinical significance of their change ("recovered," "improved," "no change" or "deteriorated") based on their pre- and postscores. Top therapists had an average recovery rate of 22.40% (21.54% improved), while bottom-ranked therapists had a recovery rate of 10.61% (17.37% improved). Conversely, bottom-ranked therapists had a 10.56% rate of deterioration, while the top ranked therapists' only had 5.20% of clients deteriorate. A chi-square analysis was computed on the clinical significance frequencies. Results of this procedure indicated that there was a significant difference between the proportion of the top- and bottom-ranked therapists' clients who fell into each category ($\chi^2_3 = 18.92$, $p < .001$), with top therapists having significantly more clients in the "recovered" and "improved" classification and significantly less in the "deteriorated" classification than the bottom-ranked therapists.

Discussion

In clinical trials, examining which therapy is most effective the individual therapist is seen as a nuisance variable that requires experimental control. Nevertheless, even in studies that attempt to minimize the individual therapist's contribution, individual therapists may contribute a relatively substantial amount of variance to client outcome (e.g., Crits-Christoph & Mintz, 1991). In applied settings and routine practice, rather than viewing differences in client outcome related to specific therapists as being problematic, such differences are viewed as providing an opportunity to enhance client outcome. More often than not, data that could be used to improve client care has rarely been used for such purposes. Orlinsky and Howard (1980), for example, classified therapists into three groups based on their client's outcome but never mentioned any strategy for applying the data. Millenson (1997) reported that "The research documenting that the wrong choice of hospital could triple a surgical patient's chance of dying was not used to improve the care of a single patient" (p. 159).

The purpose of this study was to examine the outcome of a large number of clients seen by a substantial number of therapists working in a college counseling center and to use the information to provide feedback to therapists about the relative outcome of cases with whom they worked. This was seen as a quality improvement effort that was based on the importance of the individual therapist for positive psychotherapy outcome (Brown, et al., 2004; Lambert & Okiishi, 1997; Kim, Wampold, & Bolt, 2006).

Clients in the current sample experienced statistically significant gains during treatments that lasted an average of 9.7 sessions. There were no significant differences in client outcome based on the four therapist variables of sex, level of training, type of training, or theoretical orientation, a finding that is consistent with the research literature (e.g., Beutler, Machado, & Neufeldt, 1994). Although therapists appeared to have roughly equivalent caseloads, there were significant, sometimes dramatic differences, between therapists in terms of the number of sessions they saw clients, the speed of client change, and the overall amount of change based on pre- and posttherapy change scores. Averaging these indexes of efficiency and effectiveness, the top and bottom 10% of therapists had clients whose outcome was statistically and clinically different. For example, the rate of deterioration of clients seen by the bottom-ranked therapists was double that found in clients of the top-ranked therapists.

Prior to a discussion of using the results of the present study for enhancing client outcomes, two major limitations of the current study should be noted. First and foremost, clients were not randomly allocated to therapists and there remains the possibility that differences between therapists were caused by factors that were not investigated. At the center where data were collected, case assignments were made by individuals (fulltime professionals and interns) doing intakes and were routinely assigned based on schedules that matched, and to a lesser extent, clients perceived level of pathology, difficulty, goodness of personality fit with the therapist, and the like. Most often, the intake therapist kept the client after the intake interview. These treatment assignments were made at the discretion of the intake interviewer and without regard to OQ-45 scores at intake. This method of deliberate, albeit unsystematic case assignment could pose a threat to the validity (meaning) of the findings. Without random assignment, it could be possible for a particular therapist to be given a disproportionate number of "easy" or "hard" cases, thus inflating or deflating their level of effectiveness. Case mixing on severity was originally proposed to counter-balance this problem but there were no significant differences overall in the average client-intake score among therapists, and the necessity of case mixing based on level of disturbance was nullified. Nevertheless, we cannot be sure that therapists had equivalent caseloads.

Another important limitation of this study was that outcome was measured with a single self-report scale rather than multiple assessment techniques. The use of multiple outcome measures would be likely to produce more complex results and perhaps soften the differences that were found between therapists (Hill & Lambert, 2004). Although using multiple measures and assessment methods would undoubtedly change the findings and the nature of feedback, some past research supports the generality of the single measure used in this research. Past research at the center (Lueck, 2004) found that diagnosis based on a computer interview was significantly related to OQ-45 scores such that those clients meeting criteria for multiple disorders had higher OQ-45 scores. In addition, clients with more serious disorders (e.g., mood and anxiety) had higher OQ-45 scores than clients with less-serious disorders (e.g., adjustment, and v codes such as relational problems). In addition, Beckstead et al. (2003) also found consensus between OQ-45 estimates of pathology with those reported by other measures, as well as classifications of clinically significant change. These findings offer some reassurance that the OQ-45 may be picking up pathology and its improvement that would be reported through more extensive means, and that caseloads might be equal on other indicators of disturbance such as diagnosis. Nevertheless, decisions based on a single measure must be made judiciously.

Noting these limitations, but in the interest of using the data to enhance client outcome, several interventions were considered: (a) Through administrative procedures, have clients assigned to therapists according to therapist rank order, thereby maximizing client contact with the most-effective therapists; (b) make outcome information available to clients and allow them to choose their therapist; (c) request that some therapists (e.g., #66 & #67), who had exceptionally high-deterioration rates, change their job role from providing counseling to other tasks, have their work more carefully observed, or engage in continuing education; and (d) request that therapists who have higher than average change slopes (e.g., #10, 16, 18, 32, 33), lower than average sessions, and lower than average pre- to postimprovement consider seeing clients for more sessions. Most of these proposals for use of the outcome data were impossible to implement because of prestudy agreements with the therapists to keep individual identities private. Nevertheless, it is obvious that one or a combination of the actions noted above would be in clients' best interests. The dramatic effects of making outcome data more transparent has been documented in

the popular press in the treatment of cystic fibrosis, with life expectancy increasing from 5 to 20 years after providers and treatment centers took the risk of making death rates public (Gawande, 2004). One can only hope that clinics and therapists will seriously consider the advantages of making outcome data more open.

Given the agreement in the counseling center to keep data anonymous, we decided the best way to improve client outcomes would be to give confidential individualized feedback to each therapist and ask them to reflect upon their work and problem solve on their own. A sample therapist feedback report is presented in Figure 1. As can be seen, the therapist is given information that allows comparisons on categories of effectiveness (i.e., pre- and postchange), efficiency (i.e., HLM slope), and rankings. The feedback was made available to each clinician by having the research staff deliver the blindly prepared feedback report in a sealed envelope with the research identification number on it, to the statistical consultant who then wrote the clinician's name on the envelope. Hence, none of the center research staff knew the outcomes of individual therapists, and the therapists received written feedback from a person (the statistical consultant) who did not know the results for specific therapists. In the future, feedback to therapists will be repeated on a yearly basis, possibly allowing for monitoring of changes in performance.

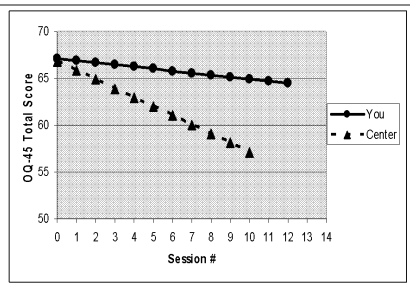
At this time it is not known if the provision of feedback to a therapist about the outcome of his or her clients in relation to that of other therapists will improve client outcome. As Saptya et al. (2005) have pointed out—feedback is most effective when it provides information on actions that can be taken to improve performance, not just information about how far away one is from hoped for goals. Because patient outcome across therapists was normally distributed, most therapists could not be distinguished from each other based on their clients' outcome. This suggests that feedback of the nature provided in the current study is only likely to help those therapists whose client outcomes were at the extreme end of the continuum.

Name: Therapist #60

Date: 1/5/05

Total # of clients: 165

	You	Center
Avg.		
Avg. # of Sessions:	12.15	9.72
HLM Slope: (avg. OQ-45 change/Session)	-.62	-.88
Slope x Average # of sessions:	-7.51	-8.14
Pre-Post Change:	-2.65	-9.43
Avg. Intake OQ-45 Score:	67.08	66.81



Summary:

This report is based on clients seen at the Center over the preceding years who were seen for three or more sessions *and* by a therapist who saw a minimum of 30 clients (You saw 165). Using these criteria, outcomes across 71 therapists were studied. Your composite ranking was 54 out of 71. 9% of your clients recovered, 17% improved, and 19% deteriorated, compared to the respective center averages of 17% recovered, 18% improved, and 8.5% deteriorated. On average your clients were seen for 12.15 sessions and their OQ-45 scores dropped 2.65 OQ-45 points during that period.

Recommendations:

Given that your patients' outcomes were in the bottom quartile, and the number who worsened was substantially higher than other therapists, it is recommended that you reflect on your work and develop a plan of action to improve the quality of care your patients are receiving.

Figure 1. Therapist feedback sheet.

A closer look at the more-effective therapists may be possible in future research, if these individuals are willing to identify themselves and be studied. In this context, volunteers from the top therapists will be sought and asked if they will permit videotape recording of their therapy sessions. Of particular interest in the current setting are therapists #1, 2, 3, and 5, whose clients showed relatively substantial change per session and substantial pre- to postchange, suggesting they are both efficient and effective. Session recordings could be analyzed for process variables that might illuminate therapeutic activities that may be related to the unusual success of their clients. The disadvantage of such traditional research is, of course, the expense, time, and slow integration of findings into training and routine practice.

The current study stands in contrast to other studies that have identified "master therapists" through processes such as peer nomination. For example, in one study of master therapists (Jennings & Skovholt, 1999), "expert" therapists (i.e., those who had written books on psychotherapy or who had been involved in training therapists) nominated colleagues: (a) to whom they would apply the term "master therapist," (b) to whom they would refer a family member or close friend, and (c) in whom they would have full confidence in seeing for their own personal therapy. While such definitions of master therapist represent one potentially valid approach to defining such a concept, we would argue that client outcome data might provide the primary basis for defining the concept of master therapist and the basis on which a referral might be best made.

We see monitoring outcomes on the basis of individual therapists as an important research activity; one that may eventually lead to the identification of the empirically validated psychotherapist, most-effective practices, and modification of theories of change. Of more practical concern is how to best use the information to improve outcomes for clients seeking treatment in our counseling center. We believe that improved outcomes are likely to come about through self-improvement activities initiated by the therapists themselves, changes in referral patterns at intake, or by client selection of effective therapists (provided outcome information can be made public). In fact, given the amount of variability among therapist outcomes in the current study and the discovery of similar findings in other investigations (Crits-Christoph & Mintz, 1991; Luborsky et al., 1985; Orlinsky & Howard, 1980; Ricks, 1974), it is arguable that the identification and study of the individual psychotherapist may be a highly effective way of improving the effects of treatment.

References

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. Journal of Personality Assessment, 82, 60–70.

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

Beckstead, D. J., Hatch, A. L., Lambert, M. J., Eggett, D. L., Goates, M. K., & Vermeersch, D. A. (2003). Clinical significance of the Outcome Questionnaire (OQ-45.2). The Behavior Analyst Today, 4, 79–90.

Beutler, L. E., Machado, P. P. P., & Neufeldt, S. A. (1994). Therapist variables. In A. E. Bergin & S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change (4th ed., pp. 229–269). New York: Wiley.

Brown, G. S., Lambert, M. J., Jones, E. R., & Minami, T. (2005). Identifying highly effective therapists in a managed care environment. American Journal of Managed Care, 11, 513–520.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Crits-Christoph, P., & Gallop, R. (2006). Therapist effects in the TDCRP and other psychotherapy studies. Psychotherapy Research, 16, 178–181.

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 University School of Medicine, Clinical Psychometrics Unit.

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

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.

Gawande, A. (2004, December). The bell shaped curve: What happens when patients find out how good their doctors really are? The New Yorker, 82–91.

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (Ed.), Bergin & Garfield's handbook of psychotherapy and behavior change (5th ed., pp. 84–135). New York: Wiley.

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.

Jennings, L., & Skovholt, T. M. (1999). The cognitive, emotional and relational characteristics of master therapists. Journal of Counseling Psychology, 46, 3–11.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. Journal of Consulting and Clinical Psychology, 67, 285–299.

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., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., et al. (2004). Administration and scoring manual for the Outcome Questionnaire-45. Orem, UT: American Professional Credentialing Services.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), Bergin and Garfield's' handbook of psychotherapy and behavior change (5th ed., pp. 139–193). New York: Wiley.

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

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

Lueck, W. C. (2004). Differential treatment response as a function of diagnostic classification. Unpublished doctoral dissertation, Brigham Young University, Provo, UT.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. Journal of Consulting and Clinical Psychology, 66, 400–410.

Millenson, M. L. (1997). Demanding medical excellence: Doctors and accountability in the information age. Chicago: University of Chicago Press.

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

Orlinsky, D. E., & Howard, K. I. (1980). Gender and psychotherapeutic outcome. In A. M. Brodsky & R. T. Hare-Mustin (Eds.), Women and psychotherapy (pp. 3–34). New York: Guilford Press.

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. 231–258). Minneapolis: University of Minnesota Press.

Sapyta, J., Riemer, M., & Bickman, L. (2005). Feedback to clinicians: Theory, research, and practice. Journal of Clinical Psychology, 61, 145–153.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multi-level models hierarchical models, and individual growth models. Journal of Behavioral and Educational Statistics, 24, 323–255.

Spielberger, C. D. (1983). Manual for the State-Trait Anxiety Inventory: STAI (Form Y). Palo Alto, CA: Consulting Psychologists Press.

Strupp, H. H. (1980). Success and failure in time-limited psychotherapy: Further evidence (Comparison 4). Archives of General Psychiatry, 37, 947–954.

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

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? Journal of Counseling Psychology, 51, 38–49.

Wampold, B. E. (2001). The great psychotherapy debate. Mahwah, NJ: Erlbaum.

Zung, W. W. (1965). A self-rating depression scale. Archives of General Psychiatry, 12, 63–70.