Journal of Consulting and Clinical Psychology

© 2014 American Psychological Association
0022-006X/14/$12.00 http://dx.doi.org/10.1037/a0036915

Benchmarking Outcomes in a Public Behavioral Health Setting: Feedback as a Quality Improvement Strategy

Robert J. Reese
University of Kentucky

Barry L. Duncan
The Heart and Soul of Change Project, Jensen Beach, Florida

Robert T. Bohanske
Southwest Behavioral Health Services, Phoenix, Arizona

Jesse J. Owen
University of Louisville

Takuya Minami
University of Massachusetts Boston

Objective: The purpose of this study was to evaluate the effectiveness of a large public behavioral health (PBH) agency serving only clients at or below the federal poverty level that had implemented continuous outcome feedback as a quality improvement strategy. ***Method:*** The authors investigated the post treatment outcomes of 5,168 individuals seeking treatment for a broad range of diagnoses who completed at least 2 psychotherapy sessions. The Outcome Rating Scale (ORS; Duncan, 2011; Miller & Duncan, 2004) was used to measure outcomes. Clients had a mean age of 36.7 years and were predominantly female (60.7%) and White (67.8%), with 17.7% being Hispanic, 9.3% being African American, and 2.8% being Native American. Forty-six percent were diagnosed with depression, mood, and anxiety disorders; 18.8% were diagnosed with substance abuse disorders; and 14.4% were diagnosed with bipolar disorder and schizophrenia. A subset of clients with a primary diagnosis of a depressive disorder was compared to treatment efficacy benchmarks derived from clinical trials of major depression. Given that the PBH agency had also implemented an outcome management system, the total sample was also compared to benchmarks derived from clinical trials of continuous outcome feedback. ***Results:*** Treatment effect sizes of psychotherapy delivered at the PBH agency were comparable to effect size estimates of clinical trials of depression and feedback. Observed effect sizes were smaller, however, when compared to feedback benchmarks that used the ORS. ***Conclusions:*** Services to the poor and disabled can be effective, and continuous outcome feedback may be a viable means both to improve outcomes and to narrow the gap between research and practice.

Keywords: effectiveness, public behavioral health, benchmarking, outcome management, feedback

Psychotherapy has demonstrated its efficacy in randomized clinical trials (RCTs; Duncan, Miller, Wampold, & Hubble, 2010; Lambert, 2013), but effectiveness in the public domain with the non-insured, poor, and those designated as disabled by Social Security is largely unknown. This is noteworthy because 61% of mental health and substance abuse care in the United States is publicly funded (Kaiser Commission on Medicaid and the Uninsured, 2011). The research findings that do exist pertaining to outcomes in public behavioral health (PBH) settings are generally not encouraging. For example, Weersing and Weisz (2002) compared outcomes for childhood depression in six community mental health centers (CMHC) in the Los Angeles area to a clinical trial benchmark derived from meta-analysis. They found the symptom trajectory of depressed youth treated in CMHCs approximated that of *control* groups in clinical trials. Perhaps the most negative data regarding services conducted in PBH settings were presented by Hansen, Lambert, and Forman (2002), who reported a 20.5% reliable change rate and 8.6% clinically significant change rate in

Robert J. Reese, Department of Educational, School, and Counseling Psychology, University of Kentucky; Barry L. Duncan, The Heart and Soul of Change Project, Jensen Beach, Florida; Robert T. Bohanske, Southwest Behavioral Health Services, Phoenix, Arizona; Jesse J. Owen, Department of Educational and Counseling Psychology, Counseling, and College Student Personnel, University of Louisville; Takuya Minami, Department of Counseling and School Psychology, University of Massachusetts Boston.

Barry L. Duncan is a co-holder of the copyright of the Outcome Rating Scale/Session Rating Scale family of measures. The measures are free for individual use at https://heartandsoulofchange.com, but Barry L. Duncan receives royalties from licenses issued to groups and organizations. In addition, there are web-based applications of the Partners for Change Outcome Management System (PCOMS), commercial products for which he receives royalties based on sales or profits from sales. We thank Ellen Usher, University of Kentucky, for her exceptional editorial feedback in the preparation of this article.

Correspondence concerning this article should be addressed to Robert J. Reese, Department of Educational, School, and Counseling Psychology, University of Kentucky, 239 Dickey Hall, Lexington, KY 40506-0017. E-mail: jeff.reese@uky.edu

a CMHC. This study seemed to confirm the conclusion reached by the President's New Freedom Commission on Mental Health (2002): "America's mental health service delivery system is in shambles [and] . . . incapable of efficiently delivering . . . effective treatments" (p. ii). Hansen et al. (2002) also reported a 35% reliable and clinically significant change rate across six different types of outpatient settings. In other words, almost two-thirds of the 6,072 clients did not report benefit from psychotherapy.

Given these findings, quality improvement strategies have garnered interest as research has moved from establishing efficacy in RCTs to demonstrating effectiveness in natural settings. A primary approach to improving the quality of mental health care is to transport evidence-based treatments into practice settings (Laska, Gurman, & Wampold, 2013; McHugh & Barlow, 2012). For example, researchers who applied evidence-based cognitive treatments for panic and depression to public behavioral health settings found that pre–post treatment effects were similar to those obtained in RCTs (Merrill, Tolbert, & Wade, 2003; Wade, Treat, & Stuart, 1998).

Some researchers have recommended that transporting evidence-based treatments should not be the only quality improvement strategy. For example, Laska et al. (2013) suggested that the utility of transporting evidence-based treatments is partially contradicted by findings from comparisons of treatment-as-usual (TAU) to benchmarks of treatments in RCTs; i.e., clients who received TAU psychotherapy in managed care and university counseling center settings likely received treatment as effective as clients receiving treatments in clinical trials (Minami et al., 2009; Minami, Wampold, et al., 2008). Practicing therapists in these settings appear to be achieving, on average, similar outcomes to RCTs, arguing against the utility and cost of transporting evidence-based treatments as a sole method to improve outcome (Laska et al., 2013).

Another strategy of quality improvement is continuous outcome feedback (Lambert, 2010). Two continuous monitoring and feedback interventions have demonstrated gains in RCTs and are included in the Substance Abuse and Mental Health Administration's National Registry of Evidence based Programs and Practices (NREPP). The first, Lambert and colleagues' Outcome Questionnaire (OQ) System, has demonstrated significant gains over TAU in six RCTs with clients at-risk for negative outcome or dropout (Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert et al., 2001, 2002; Slade, Lambert, Harmon, Smart, & Bailey, 2008; Whipple et al., 2003). A meta-analytic review of the six studies ($N = 6{,}151$) using the OQ System revealed that clients in a feedback condition had less than half the odds of experiencing deterioration and approximately 2.6 times higher odds of attaining reliable improvement than did those in a TAU condition (Shimokawa, Lambert, & Smart, 2010).

The second NREPP listed method of using continuous client feedback to improve outcomes, the Partners for Change Outcome Management System (PCOMS; Duncan, 2012, 2014; Duncan & Sparks, 2010), has demonstrated significant treatment gains for feedback over TAU in three RCTs (Anker, Duncan, & Sparks, 2009; Reese, Norsworthy, & Rowlands, 2009; Reese, Toland, Slone, & Norsworthy, 2010). Anker et al. (2009) randomized 205 couples seeking couple therapy to feedback or TAU. Compared to couples who received TAU, nearly four times as many couples in the feedback condition reached clinically significant change. Reese et al. (2009) found significant treatment gains for individual clients in the feedback condition when compared to those receiving TAU in both a university counseling center ($N = 74$) and a graduate training clinic ($N = 74$). In addition, clients in the feedback condition achieved reliable change in significantly fewer sessions than did those receiving TAU. The last RCT (Reese et al., 2010) replicated the Anker et al. study with couples and found nearly the same results ($N = 92$). In a recent meta-analysis of PCOMS studies ($N = 558$), Lambert and Shimokawa (2011) reported that clients in the feedback group had 3.5 times higher odds of experiencing reliable change and less than half the odds of experiencing deterioration.

Although promising as a quality improvement strategy, PCOMS has not been systematically evaluated in a public behavioral health setting. In addition, the findings highlighted by Laska et al. (2013) of comparable results of clinical trial treatment benchmarks and treatment as usual in university or managed care settings may or may not apply to PBH settings, given the differences in client populations. The current study, therefore, adopted a benchmarking methodology to evaluate the effectiveness of services provided to racially and ethnically diverse clients at or below the federal poverty line at a large, public behavioral health agency that implemented a continuous outcome management system, PCOMS, as a quality improvement strategy (see Bohanske & Franczak, 2010). Benchmarking permits comparison of treatments delivered in non-controlled settings against a reliably determined effect size in clinical trials or meta-analyses of clinical trials (McFall, 1996; Merrill et al., 2003; Minami, Wampold, Serlin, Kircher, & Brown, 2007; Wade et al., 1998; Weersing & Weisz, 2002).

We used the benchmarking methodology from Weersing and Weisz (2002) and Minami, Serlin, Wampold, Kircher, and Brown (2008) for the current study. Weersing and Weisz (2002) advanced previous benchmarking efforts in four ways (Minami, Serlin, et al., 2008). First, they did not alter the treatment being evaluated in the naturalistic setting which permits the results to be generalized to TAU in the same setting. Second, they used a meta-analytically determined benchmark rather than a few studies to provide a more rigorous, comprehensive comparison. Third, they included a wait-list/control condition benchmark to also compare treatment to the natural remission of symptoms. Last, they evaluated whether the effect size of interest fell within the two-tailed 95% confidence interval of the benchmark effect size rather than subjectively comparing the effect sizes. Further advancing this methodology, Minami, Serlin, et al. (2008) included the "good enough principle" (Serlin & Lapsey, 1985, 1993), which establishes a clinically relevant margin between the observed effect size and benchmark to avoid obtaining statistical significance with differences that are clinically trivial. These improved benchmarking approaches have not only enabled a better understanding of the effectiveness of clinical services in community mental health care (Weersing & Weisz, 2002), managed care (Minami, Wampold, et al., 2008) and university counseling settings (Minami et al., 2009), but have also provided a detailed methodology to allow other data sets to be similarly analyzed.

Two related questions guided our analyses. First, does continuous outcome feedback as a quality improvement strategy offer a viable alternative to the dissemination of evidence-based treatments? Second, is psychotherapy effective in a public behavioral setting serving individuals who are impoverished and/or designated as disabled? This second question arises from two findings in the literature: (a) the noted poor outcomes of services provided in PBH settings (e.g., Hansen et al., 2002) and (b) the finding that individuals from low socioeconomic backgrounds have a higher risk for psychological dysfunction and limited access to resources (e.g., Jokela, Batty, Vahtera, Elovainio, & Kivimäki, 2013; McLaughlin, Costello, Leblanc, Sampson, & Kessler, 2012; Pan, Stewart, & Chang, 2013; Reiss, 2013).

We enlisted two benchmarking strategies to address these questions. First, following the standards set by earlier studies of managed care and university counseling settings, we evaluated the effectiveness of psychotherapy provided to clients with depressive disorders at a PBH agency by comparing the observed pre–post effect size estimate against treatment efficacy benchmarks constructed from treatments in clinical trials of major depression (Minami et al., 2007). We also compared this sample of clients with depressive disorders to a benchmark of clients diagnosed with major depression who did not receive treatment. We hypothesized that the treatment offered in the PBH setting for depressive disorders would be equivalent to treatment efficacy observed in clinical trials of major depression and superior to waitlist controls. Second, given that the current PBH agency had implemented a continuous outcome management system, the overall sample was compared against benchmarks derived from clinical trials evaluating outcome feedback compared to TAU (Lambert & Shimokawa, 2011). Here we used the benchmarks derived from feedback and TAU conditions. We hypothesized that results attained in the PBH agency would be similar to benchmarks reported in RCTs of outcome feedback and superior to benchmarks derived from TAU.

Method

Participants

Participants included in this study were drawn from an archival data set of therapy outcomes at a large PBH agency, Southwest Behavioral Health Services (SBHS), a non-profit, comprehensive community behavioral health organization providing services to people living in Maricopa (Phoenix), Mohave, Yavapa, Coconino, and Gila counties in Arizona. SBHS provides clinical services to a diverse group of Medicaid insured clients at or below 100% of the federal poverty level through a wide variety of programs, including mental health and substance abuse treatments for youth and adults. The data for this study were collected from adult discharged cases between January 2007 and December 2011.

Clients (N = 5,186) were predominantly female (60.7%) and Caucasian (67.8%), ranging in age from 18 to 87 (M = 36.7, Mdn = 47.6, SD = 12.3). Most ranged in age from 18 to 40 (61.8%) or 41 to 64 (37.3%). As can be seen in Table 1, Hispanics were the largest minority (17.7%) followed by African Americans (9.3%), Native Americans (2.8%), and other ethnic groups (2.4%). Clients attended a mean of 8.86 sessions (Mdn = 5.00, SD = 10.85). Regarding primary diagnosis, depression, mood, and anxiety disorders (excluding Bipolar Disorder) were the most common (46.0%), followed by substance abuse disorders (18.8%), Bipolar Disorder and Schizophrenia (14.4%), and Adjustment Disorder (10.0%). A mix of other diagnostic categories accounted for the remainder (see Table 2 for a full list). Therapists conducted semi-structured intakes and determined a primary diagnosis by the third session. Information about comorbidity and medication use was not available.

Therapists

Therapists (N = 86) were predominantly female (84.2%) and were Caucasian (88.1%), Hispanic (9.8%), and African American (2.1%). Providers were licensed and had a master's degree or higher in one of the following fields: counseling (68.2%), clinical social work (12.7%), substance abuse counseling (11.3%), and psychology (9.4%).

The Outcome Rating Scale (ORS)

Psychological functioning and distress was assessed pre and post treatment using the Outcome Rating Scale (ORS; Duncan, 2011; Miller & Duncan, 2004), a self-report instrument designed to measure client progress repeatedly (at the beginning of each session although only first and last session data were available in the data set) throughout the course of therapy. The ORS assesses four dimensions: (1) Individual—personal or symptomatic distress or well-being, (2) Interpersonal—relational distress or how well the client is getting along in intimate relationships, (3) Social—the client's view of satisfaction with work/school and relationships outside of the home, and (4) Overall—general sense of well-being. The ORS translates these four dimensions into a visual analog format of four 10-cm lines, with instructions to place a mark on each line with low estimates to the left and high to the right. The four 10-cm lines add to a total score of 40. The score is the summation of the marks made by the client to the nearest millimeter on each of the four lines, measured by a centimeter ruler or template. Lower scores reflect more distress.

Table 1
Therapy Outcomes by Race/Ethnicity

Race/ethnicity	Pre ORS *M* (*SD*)	Post ORS *M* (*SD*)	Change *M* (*SD*)	Within-group *d* [95% CI]
Hispanic (N = 914)	20.27 (8.56)	26.34 (8.92)	6.07 (9.24)	0.71 [0.32, 1.10]
African American (N = 478)	19.53 (8.55)	25.50 (9.38)	5.97 (9.36)	0.70 [0.16, 1.24]
Native American (N = 143)	21.07 (8.40)	26.85 (9.16)	5.78 (9.41)	0.69 [−0.28, 1.66]
Asian American (N = 22)	18.91 (7.02)	25.59 (9.96)	6.68 (9.53)	0.97 [−1.05, 3.00]
Euro-American (N = 3,503)	19.01 (7.97)	24.71 (8.97)	5.71 (9.08)	0.72 [0.53, 0.90]
Other (N = 104)	21.26 (8.43)	26.79 (9.81)	5.53 (9.76)	0.66 [−0.48, 1.80]

Note. N = 5,164; 22 clients did not indicate race/ethnicity. ORS = Outcome Rating Scale; CI = confidence interval.

Table 2
Therapy Outcomes by Diagnosis (Dx)

Dx	Sample size *N*	Pre ORS *M* (*SD*)	Post ORS *M* (*SD*)	Within-group effect size *d* [95% CI]
Substance Dx	957	23.24 (8.26)	28.59 (8.55)	0.65 [0.28, 1.02]
Mood Dx NOS	633	17.59 (7.61)	24.01 (8.66)	0.84 [0.47, 1.22]
Anxiety Dx	503	19.56 (7.20)	25.15 (8.78)	0.78 [0.33, 1.22]
Schizophrenia Dx	171	19.23 (8.32)	23.75 (9.62)	0.55 [−0.33, 1.42]
Bipolar Dx	562	17.79 (8.14)	23.59 (9.05)	0.71 [0.24, 1.19]
Major depression/depression NOS	1,129	17.06 (7.84)	23.16 (9.25)	0.78 [0.46, 1.10]
Dysthymic Dx	71	19.02 (6.76)	24.95 (8.59)	0.88 [−0.22, 1.99]
Adjustment Dx	506	19.93 (7.79)	26.24 (8.29)	0.81 [0.33, 1.29]
PTSD	170	17.90 (8.36)	24.41 (9.67)	0.78 [−0.11, 1.67]
Impulse Dx	29	21.33 (8.38)	27.09 (8.39)	0.70 [−1.42, 2.82]
ADHD	73	21.84 (8.22)	25.86 (8.51)	0.49 [−0.83, 1.82]
V-codes	275	21.30 (7.22)	26.57 (8.43)	0.73 [0.13, 1.33]

Note. $d = [1 - (3/4n - 5)]\,[M\text{post} - M\text{pre}/SD\text{pre}]$. ORS = Outcome Rating Scale; CI = confidence interval; Substance Dx = any substance abuse/dependency diagnosis; NOS = not otherwise specified; Anxiety Dx = diagnosis of panic, panic with agoraphobia, anxiety NOS, phobia, obsessive-compulsive disorder, or generalized anxiety disorder; Adjustment Dx = any adjustment diagnosis; PTSD = posttraumatic stress disorder; ADHD = attention-deficit/hyperactivity disorder; V-codes = any V-code diagnosis. Diagnoses reflect the primary diagnosis. $N = 5{,}079$. There were some missing data based on diagnoses being infrequently diagnosed (e.g., learning and communication disorders, autism, and deferred diagnoses).

In addition to the PCOMS manual (Duncan, 2011; Miller & Duncan, 2004), four validation studies of the ORS have been published (Bringhurst, Watson, Miller, & Duncan, 2006; Campbell & Hemsley, 2009; Duncan, Sparks, Miller, Bohanske, & Claud, 2006; Miller, Duncan, Brown, Sparks, & Claud, 2003). Across studies, average Cronbach's alpha coefficients for ORS scores were .85 (clinical samples) and .95 (nonclinical samples; Gillaspy & Murphy, 2011). As an indicator of treatment progress, ORS scores have been found to be sensitive to change for clinical samples yet stable over time for nonclinical samples (Bringhurst et al., 2006; Duncan et al., 2006; Miller et al., 2003). The concurrent validity of ORS scores has been examined through correlations with established outcome measures. For example, the average bivariate correlation between the ORS and the OQ-45 across three studies (Bringhurst et al., 2006; Campbell & Hemsley, 2009; Miller et al., 2003) was .62 (range = .53–.74), indicating moderately strong concurrent validity (Gillaspy & Murphy, 2011).

Jacobson and Truax's (1991) formulas were used to determine the ORS clinical cutoff and the reliable change index for evaluating clinically significant change. Miller et al. (2003) used a nonclinical, community sample ($n = 86$) and a clinical sample ($n = 435$) to establish a cut score of 25.[1] The reliable change index for the ORS was computed using a diverse sample of 34,790 participants who were primarily of low socioeconomic status; the reliable change index was determined to be 5 points (Duncan, 2011; Miller & Duncan, 2004). Therefore, to achieve clinically significant change a client must begin treatment with an ORS score < 25, improve by at least 5 points, and finish treatment with an ORS score ≥ 25.

Procedures

Feedback process. SBHS implemented PCOMS beginning in 2007, eventually rolling it out across all clinical services (Bohanske & Franczak, 2010). PCOMS involves ongoing assessment of outcome using the Outcome Rating Scale (Miller et al., 2003) and the therapeutic alliance using the Session Rating Scale (SRS; Duncan et al., 2003). PCOMS is designed to identify clients who are not responding to therapy so that the lack of progress can be addressed and new approaches collaboratively developed.

Clients complete the ORS at intake and prior to each session and the SRS toward the end of each session. In the first meeting, the ORS assesses where the client sees him or herself, allowing for an ongoing comparison in later sessions. The SRS allows for routine discussion of the therapeutic alliance. The therapist and client review the client's responses on the SRS and discuss any potential alliance ruptures and how the service may be improved. At second and subsequent sessions, interpretation of the ORS depends on both the amount and rate of change that has occurred since the prior visit(s). The longer therapy continues without measurable change, the greater the likelihood of dropout and/or poor outcome. ORS scores are used to engage the client in a conversation about progress, and more important, what, if anything, should be done differently if progress is not occurring. PCOMS is designed to directly involve clients in all decisions affecting their care (Duncan, 2014).

Therapists received two days (12 hr) of PCOMS training plus annual one-day booster trainings. Although there were no fidelity checks, therapists were expected to collect outcome data, and at-risk clients identified by the data were routinely discussed in regular agency supervision. SBHS did not mandate or monitor the treatment approach used by the providers but required that they use PCOMS.

Participant inclusion criteria for depression and complete samples. The data set initially consisted of 8,224 adult clients. To answer the research questions of interest, only clients with pre and post treatment scores were included (clients must have attended at least two sessions). Given that clients at SBHS were

[1] Jacobson and Truax's (1991) cutoff formula was used: $c = (SD_0M_1 + SD_1M_0)/(SD_1 + SD_0)$; 0 = nonclinical sample, 1 = clinical sample.

asked to complete the Outcome Rating Scale at the beginning of each session, a larger inclusion rate was obtained than typical in other naturalistic data sets (e.g., Minami, Wampold, et al., 2008; Stiles, Barkham, Connell, & Mellor-Clark, 2008). However, 2,152 (26.2%) participants were eliminated because they only attended one session. Clients who were absent from services for 90 days were considered closed cases. If such clients re-entered services, only the first encounter was included. Consequently, another 293 (3.6%) clients were eliminated. Although some researchers only include clients functioning in the clinical range at intake (e.g., Minami et al., 2009; Stiles et al., 2008), we included clients whose functioning at intake was in the non-clinical range. Including this group allowed our data set to be more representative of individuals served in PBH and accounted for 27% of the final sample. Using the criteria, we identified 5,168 clients seen by 86 therapists. This complete sample was used for comparison to the feedback benchmarks.

Although the total sample is likely more representative of typical agency practice and was used to compare to the benchmarks for feedback and TAU, to address the first hypothesis and approximate the methodology used in the depression benchmarking studies of managed care and university counseling settings (Minami et al., 2009; Minami, Wampold, et al., 2008), the data were trimmed by eliminating those clients who scored over the clinical cutoff (initial score in the nonclinical range) and who had a diagnosis of any disorder other than a depressive disorder. We included clients with a primary diagnosis of major depressive disorder, dysthymia, depressive disorder not otherwise specified (NOS), and adjustment disorder with depressed mood. We did not further differentiate the depressive diagnoses for two reasons. First was the concern for the accuracy of a differential diagnosis (e.g., major depression disorder vs. depressed mood NOS) without a formalized, structured assessment process. Second and more pragmatically, the effect sizes varied little across depressive diagnoses and did not warrant further differentiation. This reduced the sample to 1,589 clients for the first benchmark comparison.

Benchmarking Strategy

Depression benchmarks. The effectiveness of treatment for SBHS clients diagnosed with a depressive disorder ($n = 1{,}589$) was evaluated by comparing it to two sets of benchmarks. The first set of benchmarks were developed using the results of RCTs focused on treating adult major depression. We selected Minami et al.'s (2007) benchmarks that provided aggregated clinical trial effect sizes derived from pre–post treatment scores for adult major depression (i.e., intent-to-treat samples [$d_{\mathrm{DEPitt}} = 0.80$] and completer samples [$d_{\mathrm{DEPc}} = 0.93$]) and waitlist control conditions for depression ($d_{\mathrm{WLC}} = 0.15$).

These three benchmarks were selected for two reasons. First, given how little is known about general effectiveness in PBH, we believed that having additional benchmarks for a common presenting issue would help contextualize our findings. Second, we selected Minami et al.'s (2007) benchmarks because the treatment efficacy studies utilized general distress outcome measures that were likely comparable in terms of sensitivity and reactivity. Given that the ORS is a general distress outcome measure, it likely has lower sensitivity and reactivity in comparison to outcome measures for an identified issue (e.g., Beck Depression Inventory or Hamilton Depression Rating Scale) that are likely to be higher on both sensitivity and reactivity resulting in higher effect sizes (Minami et al., 2007). Consistent with the adult depression treatment benchmarks, we only analyzed the pre–post data of clients who began treatment in the clinical range (ORS < 25) and were diagnosed with major depressive disorder, dysthymic disorder, depressed mood NOS, or adjustment disorder with depressed mood.

Feedback (complete sample) benchmarks. The second set of benchmarks was aggregated from RCT studies using continuous outcome feedback systems. We focused on studies that utilized the OQ System or PCOMS because the OQ System has the most research support among feedback systems, PCOMS research permitted a direct comparison to the SBHS sample, and these are the only two systems designated as evidence based. We conducted a thorough search of the peer-reviewed literature using the search terms "patient focused research," "client feedback and outcome," "OQ45," and "patient level feedback," which resulted in a total of 186 hits. We also consulted previous client feedback meta-analyses (Lambert & Shimokawa, 2011; Shimokawa et al., 2010).

Studies were excluded if they did not use a RCT design, an outpatient sample, the OQ System or PCOMS, or means and standard deviations were not provided for the entire sample. For example, Simon, Lambert, Harris, Busath, and Vazquez (2012) only provided descriptive statistics for clients who were not-on-track, and another study (Murphy, Rashleigh, & Timulak, 2012) only utilized the ORS and not the complete PCOMS method. This process resulted in nine studies selected, six for the OQ System and three for PCOMS (see Lambert & Shimokawa, 2011, for a review of the studies). To construct the feedback benchmarks, we used the formulas outlined by Minami, Serlin, et al. (2008, pp. 517–518, Formulas 1 and 2)[2] to compute unbiased, standardized effect sizes and to aggregate the effect sizes across the feedback studies (p. 518, Formula 3). We constructed four feedback benchmarks: Treatment effect sizes were calculated for the feedback ($d_{\mathrm{FTall}} = 0.60$) and TAU ($d_{\mathrm{TAUall}} = 0.41$) samples from the nine OQ/PCOMS studies and for the three PCOMS ($d_{\mathrm{FTors}} = 1.13$) and TAU ($d_{\mathrm{TAUors}} = 0.47$) samples. All clients ($N = 5{,}168$) were used in the SBHS sample irrespective of pretreatment score or diagnosis as consistent with the feedback benchmarks.

Analytical Strategy

To compare the pre–post treatment effects to the selected benchmarks, we followed the formulas and procedures highlighted in previous benchmarking studies (Minami et al., 2009; Minami, Wampold, et al., 2008; Minami et al., 2007). We used the same formula used to construct the benchmarks (Minami, Serlin, et al., 2008) to compute the observed treatment effects, $d = [1 - (3/4n - 5)]\,[M\text{post} - M\text{pre}/SD\text{pre}]$. Next, we statistically evaluated equivalence or superiority to the selected benchmarks using an a priori margin of differences between the benchmark and treatment effect sizes. Serlin and Lapsley (1985, 1993) recommend using a predetermined margin considered to be clinically trivial to resolve the dilemma of rejecting the null hypothesis with small differences

[2] Formula 2 requires the pre–post measure correlation. We used $r = .50$ for the OQ studies (Minami, Wampold, et al., 2008) and $r = .43$ for the ORS studies based on the current data set.

due to statistical power (i.e., sample size). Given our large sample size, we determined that differences between the SBHS effect sizes and the respective benchmarks that were within 10% of the benchmarks could be considered clinically negligible (Minami et al., 2009; Minami, Wampold, et al., 2008). For example, comparing against the depression intent-to-treat effect size $d_{\text{DEPitt}} = 0.80$, differences within 10% of this effect size (90%~110%, i.e., 0.72~0.88) were considered to be clinically negligible. In other words, if the treatment effect size estimate was statistically within this range or larger given a Type I error rate of $\alpha = .05$ (i.e., reject the null that the treatment effect size estimate is smaller than the lower bound of $d = 0.72$), we can conclude that the treatment effect appears to be at or above the depression intent-to-treat benchmark. Conversely, comparing against the waitlist control benchmark $d_{\text{WLC}} = 0.15$, the treatment effect size estimate must statistically exceed 110% of this benchmark (i.e., $d = 0.17$) in order to conclude that the treatment effect is larger than the waitlist control benchmark.

SBHS effect sizes were compared against the benchmarks plus the 10% margin (range-null hypotheses), which follows a noncentral *t* statistic (Serlin & Lapsley, 1985, 1993). Specifically, if the SBHS sample effect size falls at or above 90% of the treatment benchmarks (i.e., benchmark minus 10%), the SBHS effect size can be considered clinically equivalent to the benchmarks. For the comparison against the TAU and waitlist benchmarks, the 10% margin was used in the opposite direction. In other words, if the SBHS sample effect size fell within 110% of the TAU and waitlist benchmarks (i.e., benchmark plus 10%), the SBHS effect size was considered clinically equivalent to the TAU and waitlist. Therefore, to claim that the effect size estimate was superior to the TAU condition, the estimate needed to exceed 110% of the TAU and waitlist benchmarks under the specified Type I error rate.

Results

Preliminary Analyses

We screened the data for disparities in outcomes based on client gender and race/ethnicity. First, we tested whether men and women had similar outcomes via an ANOVA with ORS pre–post change scores as the DV and gender as the IV. The results for client gender were not statistically significant, $F(1, 5167) = 3.65$, $p = .06$, partial $\eta^2 = .001$. Second, we tested whether therapy outcomes varied by client race/ethnicity via an ANOVA with ORS pre–post change scores as the DV and race/ethnicity as the IV. The results for client race/ethnicity were not statistically significant, $F(5, 5158) = 0.28$, $p = .95$, partial $\eta^2 = .000$. Table 1 shows the pre and post ORS scores by race/ethnicity.

Next, we tested whether therapy outcomes varied by diagnosis, via an ANOVA with ORS pre–post change scores as the DV and primary diagnosis as the IV. The results for primary diagnosis were not statistically significant, $F(11, 5067) = 1.48$, $p = .13$, partial $\eta^2 = .003$. Table 2 shows the pre and post ORS scores by primary diagnosis. Finally, we inspected the rates of reliable and clinically significant change to provide additional context of our benchmarking results given the PBH rates reported in Hansen et al. (2002). In the total SBHS sample ($N = 5{,}168$), 65.6% achieved reliable change and 42.9% achieved clinically significant change. Table 3 presents a comparison of clinically significant change by session of the current SBHS data set and a university counseling center reported in Baldwin, Berkeljon, Atkins, Olsen, and Nielsen (2009). An inspection of Table 3 reveals a surprising similarity of the two data sets, measured by different outcome instruments (the ORS and OQ-45), in the rates of clinically significant change by session as well as the overall clinically significant rate (42.9% in the

Table 3
Clinically Significant Change by Total Number of Sessions for the SBHS Data Set and Baldwin et al. (2009)

Total *N*		*n* in clinical range		% CSC (*n*) of eligible		
SBHS	UCC	SBHS	UCC	SBHS	UCC	No. of sessions
550	N/A	420	N/A	26.2 (110)	N/A	2
702	1,195	527	706	32.8 (173)	35.8 (253)	3
549	843	401	520	38.2 (153)	40.4 (210)	4
467	597	370	381	47.3 (175)	40.4 (154)	5
360	418	251	270	43.8 (110)	42.2 (114)	6
317	311	226	208	46.9 (106)	43.3 (90)	7
280	257	186	182	51.6 (96)	46.5 (80)	8
260	229	181	153	49.7 (90)	47.7 (73)	9
213	152	155	100	51.6 (80)	50.0 (50)	10
160	128	111	92	41.4 (46)	46.7 (43)	11
144	110	101	76	54.5 (55)	47.4 (36)	12
114	93	81	60	58.0 (47)	41.7 (25)	13
107	82	68	63	45.6 (31)	49.2 (31)	14
87	43	63	32	54.0 (34)	53.1 (17)	15
91	41	63	34	50.8 (32)	47.1 (16)	16
77	32	56	23	48.2 (27)	31.1 (9)	17
586	145	435	95	49.7 (216)	43.2 (41)	18–40
104	N/A	79	N/A	46.8 (37)	N/A	41+
5,168	4,676	3,774	2,985	42.9 (1,618)	41.6 (1,242)	TOTAL

Note. SBHS = Southwest Behavioral Health Services sample; UCC = University Counseling Center sample from Baldwin et al. (2009); CSC = clinically significant change.

SBHS sample vs. 41.6% in the university counseling center sample). The mean number of sessions in the university counseling center study and the current study was 6.5 and 8.9, respectively (roughly 75% of the clients in the university counseling center study attended 8 sessions or less while approximately 75% of the SBHS sample attended 12 sessions or less). Regarding clients entering therapy in the clinical range, 63.8% of the clients in the university counseling center study entered in the clinical range compared with 72.9% in the SBHS sample.

Benchmark Comparisons

Depression benchmarks. The mean pre–post treatment ORS scores for the SBHS depressed sample ($n = 1{,}589$) were $M_{\text{pre}} = 14.73$ ($SD = 5.86$) and $M_{\text{post}} = 22.59$ ($SD = 8.86$), respectively, resulting in a standardized effect size of $d = 1.34$. This effect was statistically compared to Minami et al.'s (2007) adult major depression ITT and completer treatment efficacy benchmarks. Given the sample size, the 95th percentile critical effect size for the ITT depression benchmark ($d_{\text{DEitt}} = 0.80$) minus 10% ($d_{\text{DEitt}}[90\%] = 0.72$) was $d_{CV} = 0.76$, which was easily surpassed by the observed effect size of the treated SBHS sample (i.e., $d = 1.34$, $t = 53.42$, $df = 1{,}588$, $\lambda^3 = 28.70$, $p < .001$; see Table 4 for comparisons and critical d for each benchmark). Therefore, the pre–post treatment effect size of the SBHS subsample with depressive symptoms can be considered clinically equivalent to the pre–post treatment effect size observed in RCTs with clients who have depressive symptoms.

Compared against the completer benchmark ($d_{\text{DEc}} = 0.93$) minus 10% ($d_{\text{DEitt}}[90\%] = 0.84$), the SBHS pre–post treatment effect size was also statistically significant ($t = 53.42$, $df = 1{,}588$, $\lambda = 33.36$, $p < .001$). These findings suggest that the treatment outcomes in the SBHS sample were comparable in effectiveness to the outcomes in the clinical trials for depressed clients who completed treatment. In both cases, the SBHS effect sizes were substantially larger than the effect sizes clinical trial studies for depression.

Last, we compared the SBHS depressed sample effect size to the waitlist control benchmark effect size ($d_{\text{DWLC}} = 0.15$; plus 10% for comparison of superiority, $d_{\text{DWLC}}[110\%] = 0.17$) reported in Minami et al.'s (2007) study, which was statistically significant ($t = 53.42$, $df = 1{,}588$, $\lambda = 6.58$, $p < .001$). Again, the observed effect size was much larger than the designated benchmark. These results support the first hypothesis and suggest that treatment delivered in this PBH setting is at least comparable to treatment efficacy studies treating major depression and superior to depressed clients in a waitlist control condition.

Table 4
Effect Size Comparisons to Depression Benchmark RCT Studies

	ITT benchmark		Completers benchmark		Waitlist control benchmark	
SBHS *d*	d_{cv}	*p*	d_{cv}	*p*	d_{cv}	*p*
1.34	0.76	<.001	0.89	<.001	0.20	<.001

Note. Clients diagnosed with a depressive disorder in the Southwest Behavioral Health Services (SBHS) sample ($n = 1{,}589$) were compared to Minami et al.'s (2007) intent-to-treat (ITT) efficacy, completers, and waitlist control benchmarks. RCT = randomized clinical trial; d_{cv} = critical effect size value required to attain statistical significance.

Table 5
Effect Size Comparisons for Continuous Assessment Studies

Study	*N*	Outcome	*d*	95% CI
Current study	5,168	ORS	0.71	[0.67, 0.75]
All feedback studies	4,676	OQ-45/ORS	0.60	[0.56, 0.64]
OQ-45 studies only	4,268	OQ-45	0.57	[0.53, 0.61]
ORS studies only	408	ORS	1.13	[1.00, 1.26]

Note. Nine feedback studies were evaluated; six that utilized the Outcome Questionnaire 45.2 (OQ-45) and three that utilized the Outcome Rating Scale (ORS). CI = confidence interval.

Feedback benchmarks. The effect size estimate for the entire SBHS sample ($N = 5{,}168$), pre ORS ($M = 19.38$, $SD = 8.17$) and post ORS ($M = 25.18$, $SD = 9.05$), was $d_{\text{SBHS}} = 0.71$. The Feedback treatment benchmark using all nine RCT studies had an overall estimated effect size of $d_{\text{FTall}} = 0.60$, but there was substantial variability in effect sizes depending upon the feedback procedure and measure used as can be observed in Table 5. Therefore, we also constructed benchmarks using only the three PCOMS studies to provide a more direct comparison. To evaluate equivalence, the SBHS sample was compared to the Feedback treatment benchmark effect size of all nine RCTs minus 10% ($d_{\text{FT}}[90\%] = 0.54$). The observed sample effect size exceeded the critical $d = 0.56$ and yielded a statistically significant difference ($t = 51.04$, $df = 5{,}167$, $\lambda = 38.82$, $p < .001$), suggesting that treatment provided at SBHS was at least equivalent to the treatment conditions in the nine feedback studies. However, the observed sample effect size fell short of the benchmark margin when compared to the PCOMS Feedback treatment benchmark minus 10% ($d_{\text{FTORS}}[90\%] = 1.02$), $t = 51.04$, $df = 5{,}167$, $\lambda = 73.11$, $p > .999$). This finding suggests that the treatment received in the SBHS sample was not equivalent to the benchmark and did not achieve the standard of treatment found in the three PCOMS studies.

We also compared the SBHS sample to the TAU conditions from the feedback studies. Specifically, we compared the TAU benchmark for all nine feedback sample effect sizes plus 10% ($d_{\text{TAUall}}[110\%] = 0.45$) to the SBHS sample and found that the SBHS effect size was superior ($t = 51.04$, $df = 5{,}167$, $\lambda = 32.42$, $p < .001$). For the TAU benchmark from the three PCOMS studies ($d_{\text{TAUors}}[110\%] = 0.52$), we also found that the SBHS effect size was significantly larger ($t = 51.04$, $df = 5{,}167$, $\lambda = 37.17$, $p < .001$). Table 6 shows the critical d required to obtain statistical significance for each of the comparisons. These results partially support hypothesis two and suggest that treatment delivered at SBHS was comparable to RCT feedback studies overall and superior to TAU from the same RCT feedback studies as well as PCOMS TAU, but not of RCTs of the PCOMS Feedback condition.

Discussion

Given that the majority of mental health and substance abuse services occur in the public sector, there has been a surprising lack

[3] λ = noncentrality parameter used to estimate critical t.

Table 6
Effect Size Comparisons to Feedback Benchmark RCT Studies

	Feedback benchmark (all)		Feedback benchmark (ORS)		TAU benchmark (all)		TAU benchmark (ORS)	
d	d_{cv}	*p*	d_{cv}	*p*	d_{cv}	*p*	d_{cv}	*p*
0.71	0.56	<.001	1.05	.999	0.39	<.001	0.45	<.001

Note. RCT = randomized clinical trial; ORS = Outcome Rating Scale; TAU = treatment-as-usual; d_{cv} = critical effect size value required to attain statistical significance.

of information available about the effectiveness of psychotherapy conducted in these settings. To our knowledge this is the first benchmarking study of treatment outcomes at a PBH agency not limited to the transportation of an evidence-based treatment provided by a limited number of therapists to clients with a specific diagnosis. One of the goals of this study was to evaluate how a public behavioral system of care fared using one quality improvement strategy, continuous outcome feedback, by comparing it to: standards generated by another quality improvement strategy, benchmarks determined from meta-analysis of clinical trials; and standards arising from the quality improvement strategy employed in the current study, continuous outcome feedback, to benchmarks generated from both OQ System and PCOMS feedback studies. The use of continuous outcome feedback in the public health agency was shown to meet the standards of both strategies.

A comparison of effect size estimates revealed that psychotherapy for adult depression provided in a particular PBH setting is likely effective; providers in this study generated effect size estimates that were similar to those observed in treatment in clinical trials for major depression. In addition, the total sample effect size estimates of the PBH agency were also comparable to RCTs evaluating systematic client feedback (OQ System and PCOMS combined) but not to RCTs of PCOMS alone.

Preliminary analyses were conducted on client demographic variables such as race/ethnicity, gender, and diagnoses. The current study found that demographic variables had little impact on effectiveness. An interesting "non-finding" was that diagnosis had little impact on differential outcome as well. This should be interpreted with caution; however, given the diagnoses were based on unsystematic clinical interviews rather than structured diagnostic interviews.

Comparisons to the two noted benchmarking studies (Minami et al., 2009; Minami, Wampold, et al., 2008) of clients in the clinical range, specific to depression, revealed similar effect size estimates. The similarity is noteworthy given the representative nature of the current sample. Both of the other benchmarking studies lost considerable portions of data. For example, Minami, Wampold, et al. (2008)—who conducted the study in a managed care setting in which the OQ was administered by 65% of therapists and only required at the first, third, fifth, and every fifth session thereafter—lost over 55% of the data for lack of two data points. Recall, however, that 26% of the current data set was lost because of the attrition of the first to the second session.

Evaluation of the observed effect size estimates of depressed clients compared to the depressed client waitlist control benchmark suggested that approximately 87% of the clients treated for two or more sessions at this agency were likely better off after receiving treatment than is the average client randomized into a waitlist control condition. Therefore, despite differences in clinical and demographic characteristics between the agency and clinical trials included in the benchmark, it is reasonable to conclude that psychotherapy services that include continuous outcome feedback provided at this agency are effective.

Very few studies have systematically investigated large naturalistic data sets. We were able to identify only two other studies in addition to the benchmarking studies discussed above. Table 3 presents the comparison of the rates of clinically significant change of the current data set to those reported by Baldwin et al. (2009). The prevailing assumptions regarding the two sites may be that university counseling clients are likely to be more functional than PBH clients (more available resources, education, etc.) and therefore more likely to achieve better outcomes. While there is some support for the first assumption given the percentage of clients entering in the clinical range, the difference (9.1%) may be less than expected. The second assumption was not borne out by this study.

In the second study, even more impressive results were reported from a large U.K. sample (N = 9,703). Stiles et al. (2008) found a reliable change rate of 81.4% and a clinically significant change rate of 62% but given that they included only completers and those who had planned terminations prevent meaningful comparisons—only 9,703 clients were included from a data base of over 33,000 (Stiles et al., 2008).

The current study demonstrated outcomes superior to previous reports of outcomes in PBH settings (Hansen et al., 2002; Weersing & Weisz, 2002) and largely comparable to estimates of both benchmarks for major depression and overall feedback. Perhaps the most obvious explanation is the dose of treatment, the issue highlighted by Hansen et al. (2002), who argued that the dose of treatment (4.1 sessions in the State CMHC sample and 4.3 overall) was inadequate exposure to psychotherapy for improvement to occur. The current study provides some support for their argument given that the average was 8.9 sessions. Not supportive of the dose explanation, however, is that in as few as three sessions in the current sample (see Table 3), over 50% of clients achieved either reliable (21.7%) or clinically significant change (32.8%).

The addition of continuous client feedback as a quality improvement strategy provides a more likely explanation. Considering both the OQ System and PCOMS, identifying clients at risk via the routine use of outcome measures has now been shown in nine RCTs to improve outcomes. Southwest Behavioral Health Services started implementation of continuous client feedback in 2007, and now integrates PCOMS in all services (Bohanske & Franczak, 2010). Although not addressed as a quality improvement strategy, the managed care and university counseling center benchmark studies (Minami et al., 2009; Minami, Wampold, et al., 2008) as well as Baldwin et al.'s (2009) study discussed above also routinely monitored outcomes with the OQ. The reported comparable results to the clinical trial treatment benchmarks could be argued to be partially due to continuous outcome monitoring. This is of course an empirical question; we only are speculating given the absence of any direct comparison of continuous outcome monitoring to either TAU or a transported evidence based treatment. Future investigation could conduct such comparisons as well as quasi-experimental or cluster randomization research in the implementation of outcome feedback in other PBH settings.

A limitation of the current study is the use of one, brief outcome measure, the ORS. The ORS is by design brief and therefore feasible for routine clinical use. Its feasibility, however, is also a drawback. Although psychometrically acceptable, it does not yield the breadth or depth of information found in longer measures like the OQ-45. A major question highlighted by this study is the difference of the effect sizes of the ORS and OQ-45 found in RCTs. There are at least three possible explanations. First, the higher effect sizes of the ORS may be an artifact of the ORS itself. It may be more sensitive to or over-represent change compared to the OQ-45. Although the ORS and OQ-45 are moderately correlated and seem to result in similar expected treatment trajectories as well as similar clinically significant change rates in the comparison noted above, the sensitivity differences between the two measures may differ. This is currently being empirically investigated. Second, given that the ORS is administered and discussed with clients, it may via demand characteristics, lead to inflated scores. Follow-up results in Anker et al.'s (2009) trial included client ratings of the ORS administered via mail 6 months post treatment. The feedback effect was maintained which suggests that demand characteristics were not responsible. Finally, the effect sizes of the ORS may be related to the differences in clinical processes associated with the two measures. The PCOMS process of discussion of both outcome and the alliance at every session may explain the larger treatment gains (Anker et al., 2009; Duncan, 2012). Partial support of this possibility is provided by OQ System feedback studies in which a higher effect size occurs when clinical support tools, like alliance measures, are incorporated with the OQ-45 (Shimokawa et al., 2010). This too is an empirical question and further research will hopefully shed light here as well.

A related issue arises from our comparison of the current total sample and the PCOMS Feedback condition of the RCTs. Although equivalent to the combined OQ System and PCOMS treatment conditions and superior to the combined TAU conditions, the current total sample was not equivalent to the PCOMS Feedback benchmark. Although we again cannot definitively explain this finding, the most obvious possibility is the differences in the samples. The RCTs were conducted in settings with clients who likely had more available resources and problems of less severity and chronicity. We also are uncertain of the adherence to the PCOMS protocol by the Southwest Behavioral Health Services therapists; it is possible that therapists in the RCTs were more compliant because they knew PCOMS was being evaluated.

The limitations of benchmarking detailed by Minami, Wampold, et al. (2008) are applicable here and also call for caution in interpreting the results: (1) although the benchmarks constructed from the feedback studies partially included the same outcome measure (ORS), and although the ORS matches the low reactivity/low specificity of the benchmark instruments in general, the efficacy and natural history benchmarks taken from Minami et al. (2007) were constructed from different measures; (2) comparison against clinical trials are not ideal given that treatment provided in a PBH setting is drastically different than RCTs—no random assignment, set amount of sessions, or control of diagnostic validity, co-morbidity, or therapeutic environment (although the feedback RCT studies were likely more comparable to treatment in a PBH setting with the exception of random assignment); (3) the characteristics of the psychotherapy delivered by the therapists in this study are unknown including their orientation or use of evidence-based treatments; (4) the context of a PBH setting and use of feedback limits the generalizability to other settings, and the extent to which there was therapist fidelity to the feedback intervention is unknown; (5) therapist effects were not modeled in this study or the clinical trials to set the benchmark except for Anker et al.'s (2009) and Reese et al.'s (2010) studies; and (6) benchmarking cannot explain why clinical trials and natural settings are similar or different given the profound differences in client and therapist factors. As Minami, Wampold, et al. (2008) concluded so do we, that although not perfect, and given that there no benchmarks for effectiveness in PBH settings, benchmarks constructed from efficacy in clinical trials are the best currently available and provide some preliminary evidence of effectiveness in public settings.

Another limitation of the current study is that data on medication use were unavailable for this study. Agency estimates, however, suggest that approximately 35% of clients were on psychotropic medication. Given that Minami, Wampold, et al. (2008) reported an increase in effect ($d = 0.15$) by use of psychotropic medication in a managed care setting (although severity was not controlled), it is possible that the agency's observed effect size calculated with only depressed clients who were not on medication could have been as low as $d = 1.19$ (which is still above the ITT and completer benchmarks). Replication with medication information is therefore necessary.

In light of the limitations, the current study provides preliminary evidence that one PBH agency using a continuous client feedback system has routinely been providing effective psychotherapy services. Given previous studies of PBH settings and the serious concerns raised by the President's New Freedom Commission on Mental Health (Hansen et al., 2002; Weersing & Weisz, 2002), this study offers a tentative empirical counter to those concerns.

Perhaps more importantly, our results also suggest that adding routine outcome management as a quality improvement strategy may be a viable alternative to transporting evidence-based treatments to natural settings. Bohanske and Franczak (2010) make a similar argument based on what they called "efficiency variables." They reviewed PCOMS at several public agencies and reported substantial improvements in client retention, therapist productivity, and length of stay.

Laska et al. (2013) call for an integrated quality improvement strategy consisting of the dissemination of evidence-based treatments and a "common factors" approach which includes outcome feedback. Continuous feedback can be easily integrated with any quality improvement strategy including evidence-based treatments. The OQ System and PCOMS are not specific treatment approaches for particular diagnoses and instead are a-theoretical and can be applied to clients of all diagnostic categories (Duncan, 2012). Both feedback systems are congruent with the American Psychological Association Presidential Task Force on Evidence-Based Practice (2006; Duncan & Reese, 2012) definition of evidence-based practice in psychology. Continuous outcome feedback enables the identification of clients who are not benefiting from any given treatment so that clinicians may collaboratively design different interventions. This approach to clinical practice does not prioritize evidence-based treatments as a quality improvement strategy. Rather, it calls for a more sophisticated and empirically informed clinician who chooses from a variety of orientations and methods to best fit client preferences and cultural values. Although there has not been convincing evidence for differential efficacy among approaches (Duncan et al., 2010), there is indeed differential effectiveness for a particular approach with a particular client — therapists need expertise in a broad range of intervention options, including evidence- based treatments, but client measurable response to treatment must be the ultimate goal.

Psychotherapy science continues to move from a sole focus on the RCT and efficacy to the study of effectiveness in clinical practice. Benchmarking studies have provided the methodology to further narrow the split between research and practice. Wolfe (2012) suggests that feasible outcome tools for everyday clinical practice, like the OQ and ORS, can serve to build the bridge between research and practice. We hope that our study offers a demonstration of this possibility and encourages other looks at practice in natural settings enabled by routine outcome management (Lambert, 2010). Everyday data collection could allow many research possibilities. One possibility may be a more routine use of RCT methodology in PBH settings as well as the examination of clients of diverse ethnicity and race. Current feedback studies are unfortunately quite restricted in this area. Another compelling research question yet to be addressed is why feedback results in improved outcomes. Routine outcome management using the OQ System or PCOMS could enable dismantling strategies to address this important topic. Finally, continuous outcome monitoring could allow an ongoing evaluation of quality improvement strategies.

On October 31, 1963, President John F. Kennedy (JFK) signed into law the Community Mental Health Act (also known as the Mental Retardation and Community Mental Health Centers Construction Act of 1963). It was the last piece of legislation JFK signed before his assassination. For millions of Americans, JFK's final legislation opened the door to a new era of hope and recovery—to a life in the community. With the 50th anniversary of the Community Mental Health Act of 1963 passed, this study presents a preliminary but more hopeful picture of outcomes in PBH. While replications are necessary, our results are reassuring to those who receive, provide, or pay for services in the public sector, suggesting that therapists in a PBH setting, when given systematic outcome feedback, are effectively treating not only depression but also a range of psychological problems.

As outcome measures become more readily available to frontline practitioners and PBH agencies, a more accurate picture will likely emerge about the effectiveness of psychotherapy with those who arguably need the services most. Routine collection of outcome data, providing individualized, responsive services, and involving consumers in decisions about their care holds promise to not only inform us about the effectiveness of PBH care and the classic question of what works for whom, but also a viable strategy to ensure quality to those who are often not considered in discussions of psychotherapy.

References

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61,* 271–285. doi:10.1037/0003-066X.61.4.271

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77,* 693–704. doi:10.1037/a0016062

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

Bohanske, R., & Franczak, M. (2010). Transforming public behavioral health care: A case example of consumer directed services, recovery, and the common factors. In B. Duncan, S. Miller, B. Wampold, & M. Hubble (Eds.), *The heart and soul of change: Delivering what works* (2nd ed., pp. 299–322). doi:10.1037/12075-010

Bringhurst, D. L., Watson, C. W., Miller, S. D., & Duncan, B. L. (2006). The reliability and validity of the Outcome Rating Scale: A replication study of a brief clinical measure. *Journal of Brief Therapy, 5,* 23–30.

Campbell, A., & Hemsley, S. (2009). Outcome Rating Scale and Session Rating Scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist, 13,* 1–9. doi:10.1080/13284200802676391

Duncan, B. (2011). *The Partners for Change Outcome Management System (PCOMS): Administration, scoring, interpreting update for the Outcome and Session Ratings Scale*. Jensen Beach, FL: Author.

Duncan, B. (2012). The Partners for Change Outcome Management System (PCOMS): The Heart and Soul of Change Project. *Canadian Psychology/Psychologie canadienne, 53,* 93–104. doi:10.1037/a0027762

Duncan, B. (2014). *On becoming a better therapist: Evidence-based practice one client at a time* (2nd ed.). Washington, DC: American Psychological Association.

Duncan, B., Miller, S., Sparks, J., Claud, D., Reynolds, L., Brown, J., & Johnson, L. (2003). The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of Brief Therapy, 3,* 3–12.

Duncan, B., Miller, S., Wampold, B., & Hubble, M. (Eds.). (2010). *The heart and soul of change: Delivering what works in therapy* (2nd ed.). Washington, DC: American Psychological Association. doi:10.1037/12075-000

Duncan, B. L., & Reese, R. J. (2012). Empirically supported treatments, evidence based treatments, and evidence based practice. In G. Stricker & T. Widiger (Eds.), *Handbook of psychology: Clinical psychology* (2nd ed., pp. 977–1023). doi:10.1002/9781118133880.hop208021

Duncan, B., & Sparks, J. (2010). *Heroic clients, heroic agencies: Partners for change* (2nd ed.). Jensen Beach, FL: Author.

Duncan, B., Sparks, J., Miller, S., Bohanske, R., & Claud, D. (2006). Giving youth a voice: A preliminary study of the reliability and validity of a brief outcome measure for children, adolescents, and caretakers. *Journal of Brief Therapy, 5,* 71–87.

Gillaspy, J. A., & Murphy, J. J. (2011). The use of ultra-brief client feedback tools in SFBT. In C. W. Franklin, T. Trepper, E. McCollum, & W. Gingerich (Eds.), *Solution-focused brief therapy* (pp. 73–94). doi:10.1093/acprof:oso/9780195385724.003.0034

Hansen, N., Lambert, M., & Forman, E. (2002). The psychotherapy dose-effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9,* 329–343. doi:10.1093/clipsy.9.3.329

Harmon, S. C., Lambert, M. J., Smart, D. W., Hawkins, E. J., Nielson, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist/client feedback and clinical support tools. *Psychotherapy Research, 17,* 379–392. doi:10.1080/10503300600702331

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K., & Tuttle, K. (2004). The effects of providing patient progress information to therapists and patients. *Psychotherapy Research, 14,* 308–327. doi:10.1093/ptr/kph027

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. doi:10.1037/0022-006X.59.1.12

Jokela, M., Batty, G. D., Vahtera, J., Elovainio, M., & Kivimäki, M. (2013). Socioeconomic inequalities in common mental disorders and psychotherapy treatment in the UK between 1991 and 2009. *The British Journal of Psychiatry, 202,* 115–120. doi:10.1192/bjp.bp.111.098863

Kaiser Commission on Medicaid and the Uninsured. (2011). *Mental health financing in the United States: A primer*. Menlo Park, CA: The Henry J. Kaiser Family Foundation. Retrieved from http://www.kff.org

Lambert, M. J. (2010). "Yes, it is time for clinicians to monitor treatment outcome". In B. L. Duncan, S. C. Miller, B. E. Wampold, & M. A. Hubble (Eds.), *Heart and soul of change: Delivering what works in therapy* (2nd ed., pp. 239–266). Washington, DC: American Psychological Association.

Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 169–218). Hoboken, NJ: Wiley.

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy, 48,* 72–79. doi:10.1037/a0022238

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11,* 49–68. doi:10.1080/713663852

Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielson, S. L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy, 9,* 91–103. doi:10.1002/cpp.324

Laska, K. M., Gurman, A. S., & Wampold, B. E. (2013, December 30). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy*. Advance online publication. doi:10.1037/a0034332

McFall, R. M. (1996). Making psychology incorruptible. *Applied and Preventive Psychology, 5,* 9–15. doi:10.1016/S0962-1849(96)80021-7

McHugh, K. R., & Barlow, D. H. (2012). *Dissemination and implementation of evidence-based psychological interventions*. New York, NY: Oxford University Press.

McLaughlin, K. A., Costello, E. J., Leblanc, W., Sampson, N. A., & Kessler, R. C. (2012). Socioeconomic status and adolescent mental disorders. *American Journal of Public Health, 102,* 1742–1750. doi:10.2105/AJPH.2011.300477

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology, 71,* 404–409. doi:10.1037/0022-006X.71.2.404

Miller, S. D., & Duncan, B. L. (2004). *The Outcome and Session Rating Scales: Administration and scoring manual*. Jensen Beach, FL: Authors.

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J., & Claud, D. (2003). The Outcome Rating Scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy, 2,* 91–100.

Minami, T., Davies, D. R., Tierney, S. C., Bettmann, J. E., McAward, S. M., Averill, L. A., . . . Wampold, B. E. (2009). Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *Journal of Counseling Psychology, 56,* 309–320. doi:10.1037/a0015398

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality & Quantity, 42,* 513–525. doi:10.1007/s11135-006-9057-z

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology, 76,* 116–124. doi:10.1037/0022-006X.76.1.116

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

Murphy, K. P., Rashleigh, C. M., & Timulak, L. (2012). The relationship between progress feedback and therapeutic outcome in student counselling: A randomised control trial. *Counselling Psychology Quarterly, 25,* 1–18. doi:10.1080/09515070.2012.662349

Pan, Y. J., Stewart, R., & Chang, C. K. (2013). Socioeconomic disadvantage, mental disorders and risk of 12-month suicide ideation and attempt in the National Comorbidity Survey Replication (NCS-R) in US. *Social Psychiatry and Psychiatric Epidemiology, 48,* 71–79. doi:10.1007/s00127-012-0591-9

President's New Freedom Commission on Mental Health. (2002). *Interim report* (DHHS Pub. No. SMA-03-3932). Retrieved from http://www.mentalhealthcommission.gov/reports/interim_toc.htm

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy: Theory, Research, Practice, Training, 46,* 418–431. doi:10.1037/a0017901

Reese, R. J., Toland, M. D., Slone, N. C., & Norsworthy, L. A. (2010). Effect of client feedback on couple psychotherapy outcomes. *Psychotherapy: Theory, Research, Practice, Training, 47,* 616–630. doi:10.1037/a0021182

Reiss, F. (2013). Socioeconomic inequalities and mental health problems in children and adolescents: A systematic review. *Social Science & Medicine, 90,* 24–31. doi:10.1016/j.socscimed.2013.04.026

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83. doi:10.1037/0003-066X.40.1.73

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shimokawa, K., Lambert, M., & Smart, D. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78,* 298–311. doi:10.1037/a0019247

Simon, W., Lambert, M. J., Harris, M. W., Busath, G., & Vazquez, A. (2012). Providing patient progress information and clinical support tools to therapists: Effects on patients at risk of treatment failure. *Psychotherapy Research, 22,* 638–647. doi:10.1080/10503307.2012.698918

Slade, K., Lambert, M. J., Harmon, S., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy, 15,* 287–303. doi:10.1002/cpp.594

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

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology, 66,* 231–239. doi:10.1037/0022-006X.66.2.231

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70,* 299–310. doi:10.1037/0022-006X.70.2.299

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50,* 59–68. doi:10.1037/0022-0167.50.1.59

Wolfe, B. E. (2012). Healing the research-practice split: Let's start with me. *Psychotherapy, 49,* 101–108. doi:10.1037/a0027114

Received March 25, 2013
Revision received March 28, 2014
Accepted April 8, 2014 ■