Psychotherapy
2015, Vol. 52, No. 4, 412–421

0033-3204/15/$12.00 http://dx.doi.org/10.1037/pst0000033

A Collaborative Outcome Resource Network (ACORN): Tools for Increasing the Value of Psychotherapy

George S. (Jeb) Brown, Ashley Simon, and Joanne Cameron
Center for Clinical Informatics, Salt Lake City, Utah

Takuya Minami
University of Massachusetts, Boston

The authors describe a collaborative outcomes resource network (ACORN) and the suite of measurement and decision support tools (ACORN Toolkit) that have emerged from this collaboration for the purpose of providing clinical feedback to therapists. The ACORN Toolkit is most accurately described as a comprehensive clinical information system designed to increase the value of mental health services across large systems of care. It was built to integrate large datasets from multiple sources including outcome data, client demographics and diagnostic data, therapist credentialing information, pharmacy data, and service claims data. For the limited purposes of this article, the authors focus on the ACORN Toolkit for measuring and how it has contributed to improving outcomes in psychotherapy. Implications to current practice and future training are provided.

Keywords: outcomes measurement, feedback, case mix adjustment, alliance

The ACORN collaboration is supported by multiple third party payers, large group practices, hospitals, and nonprofit agencies. The activities of the collaboration are coordinated and facilitated by the Center for Clinical Informatics (CCI), a consulting group that includes the authors. Through decades of collaboration, the current ACORN Toolkit is now comprised of a large database of psychotherapy treatment outcomes, a clinical information system that analyzes the data on a continuous basis, and a secure web interface for users to access the information (www.psychooutcomes.org). Any user can access the Toolkit from any device capable of running a browser with an Internet connection. Here, users could be therapists, supervisors, or administrators, and, depending on their level, they would have access to different information. For example, an individual therapist is only able to view data for his or her cases, although a clinical supervisor has access to data for all clients treated by therapists under his or her supervision.

History of Research and Development Leading to ACORN

The ACORN collaboration was not built overnight, and has, over the years, involved numerous collaborators including academic researchers who have helped bridge the gap between research and practice. Its origin extends back to 1994 when the first author (George S. [Jeb] Brown) was serving as the Director of Clinical Programs for Human Affairs International, then a subsidiary of Aetna Health Plans. With funding from Aetna, Brown headed a team of researchers and software programmers tasked with developing a clinical information system to measure psychotherapy outcomes. Among the initial collaborators were academics including Michael Lambert and Gary Burlingame from Brigham Young University (BYU). Between 1994 and 1998, the team implemented a program to measure treatment outcomes using the Outcome Questionnaire-45 (OQ-45; Lambert & Brown, 1996) and the Youth Outcome Questionnaire-64 (YOQ-64; Wells, Burlingame, Lambert, Hoag, & Hope, 1996). As part of this project, Brown and Lambert began to develop algorithms to identify clients who might be at risk for early termination and poor treatment outcomes (Lambert & Brown, 1996; Brown & Lambert, 1998).

In 1998, CCI was founded as a consulting group with a contract to develop a clinical information system for PacifiCare Behavioral Health (PBH). CCI continued to collaborate with academics including Lambert and Burlingame. Bruce Wampold from the University of Wisconsin-Madison subsequently joined in 2002, as did his then graduate student, Minami, who is one of the coauthors. The clinical information system developed for PBH was named as the ALERT System (Algorithms for Effective Reporting and Treatment; Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001). The use of clinical algorithms to identify at risk clients was further informed by clinical trials on measurement and feedback conducted by Lambert and colleagues at the Comprehensive Counseling Center at BYU (for details, see Lambert, 2010a).

Therapists within the PBH network were invited to participate in the ALERT System by administering the outcome questionnaires at regular intervals in treatment, and faxing completed forms to PBH for data entry. In exchange for providing a completed outcome questionnaire, therapists were automatically authorized for

George S. (Jeb) Brown, Ashley Simon, and Joanne Cameron, Center for Clinical Informatics, Salt Lake City, Utah; Takuya Minami, College of Education and Human Development, University of Massachusetts, Boston.

Takuya Minami maintains a consulting relationship with the Center for Clinical Informatics. This research was supported by the participants in the ACORN collaboration, including six third-party payers who help fund the collaboration.

Correspondence concerning this article should be addressed to George S. (Jeb) Brown, Center for Clinical Informatics, 1821 Meadowmoor Road, Salt Lake City, UT 84117. E-mail: jebbrown@clinical-informatics.com

additional sessions with their clients with whom they provided data on. Based on these data, PBH was able to preferentially refer clients to practices with the strongest evidence of effectiveness. Increasing referrals to effective therapists improved overall outcome while managing costs, as better outcomes were associated with lower cost of care (Brown, Lambert, Jones, & Minami, 2005). By this time, the ALERT system was expanded into a true data warehouse, combining data from multiple sources including insurance claims for services provided, pharmacy claims, and credentialing information for therapists, allowing for analyses of outcomes from multiple angles.

Collaboration with researchers also yielded several methodological developments, including investigations of outcome variance attributable to therapists and its implications (Brown et al., 2005). Most notably, Wampold and Brown (2005) published one of the earliest articles quantifying the percentage of variance due to the therapist as well as the interaction between the therapists' effectiveness and effect of medication. Although it is beyond the scope of this article to discuss findings on sources of variance in treatment outcomes, it is now well known that therapists account for a far greater percentage of variance in outcomes than the method of therapy. Wampold and Imel (2015), in their comprehensive review of psychotherapy research, conclude that between 3% and 7% of the variance in psychotherapy outcomes is due to the therapist, while at most 1% is due to the method of therapy. In addition, using data from the ALERT System, Minami and colleagues advanced a methodology for benchmarking, that is, evaluating treatment outcomes in naturalistic settings against data from clinical trials. This resulted in a series of articles describing the methodology and demonstrating that the treatment outcome effect size observed in the ALERT System was clinically comparable to those observed in clinical trials (Minami et al., 2007; Minami et al., 2008a; Minami et al., 2008b).

PBH was acquired by United Health Care in 2005, and the ALERT System continued as United Health Care's primary outcomes management system. During this period, the ALERT System was updated based on the above methodological advancements in estimation of therapists' effectiveness and benchmarking. First, following studies on therapist effects, the ALERT System incorporated multilevel modeling to account for the nesting of data. At the client level, the severity adjusted effect size (SAES) adjusted the raw pre-post effect size for differences in case mix, using the severity at intake (intake score) and diagnostic cluster as predictors, which will be described in detail below. At the therapist level, effectiveness was estimated using a random effects model as a means to appropriately adjust for variability among therapists in their number of cases (Minami, Brown, McCulloch, & Bolstrom, 2012). In the ALERT System, therapists' modal SAES was approximately $d = 0.8$ with a distribution that visually resembled the normal curve. Consistent with earlier studies published using the data from the ALERT System, significant variability among therapists was observed, including a small number of therapists whose SAES was below $d = 0.5$, implying that their aggregate effectiveness was closer to clinical trials outcomes observed among clients in wait list controls than those being provided treatment. Details on the calculation of the SAES are provided in a later section, as the current ACORN System is an outgrowth of the ALERT system.

ACORN Collaboration

In 2007, CCI secured funding from several payers to develop a new clinical information system capable of supporting outcomes measurement and feedback using a wide variety of outcome measures. Rather than being developed as a proprietary system used by one payer for their closed panel of therapists, the new system would be a shared platform used by a wide variety of payers and therapists using a variety of questionnaires. The collaboration was branded as ACORN (a collaborative outcomes resource network), and a wiki site devoted to this collaboration was launched in 2007 (www.psychoutcomes.org).

Measurement 2.0 and Questionnaire Development

In 2007, Warren Lambert from Vanderbilt University, given his extensive experience in psychometrics and questionnaire development, joined the collaboration. Lambert guided the initiative to develop a large set of items for use by the participants of the ACORN collaboration, which included Regence BlueCross/BlueShield, United Health Care, ValueOptions, APS HealthCare, and Kaiser Permanente. The idea was to create an item pool, with extensive normative data for each item. As a part of this effort, various item formats were tested and decisions were made based on psychometric properties and feasibility for routine use. First, items assessing most frequently observed clinical symptoms were created. Subsequently, the items were psychometrically analyzed using methods based on both classical test theory and item response theory to ensure that they were valid, reliable, and sensitive to change. Lastly, different versions of the questionnaires were compiled by combining the items based on the specific needs of the organization or the agency. Currently, as part of the ACORN collaboration, normative data from real world clinical settings are available for over 390 items obtained from over 630,000 clients, most with multiple assessments over time. The specifications are available on the ACORN wiki,[1] but we provide the summary for our most frequently used questionnaire in the subsequent section.

Further, to capitalize on the evidence of a strong working alliance contributing to psychotherapy outcomes (Norcross & Lambert, 2011), the ACORN collaboration made attempts early on to introduce a separate alliance questionnaire at the end of the session, only to result in lackluster participation. Based on a clever, retrospectively obvious suggestion made by one of the participating therapists, we explored adding a few alliance items at the end of the outcomes questionnaires asking the clients to look back at their last session and rate their therapeutic alliance. This proved highly effective in encouraging clients to complete alliance items and had an immediate and apparent effect of reducing premature drop out and improving outcomes, which finding we will elaborate on later.

Psychometric Properties of Adult Version 11

In general, therapists and patients require brevity whereas researchers pursue comprehensiveness intended toward psychometric rigor. The ACORN collaboration balances these demands by

[1] http://psychoutcomes.org/pub/OutcomesMeasurement/OutcomesMeasurement20/Outcomes_Measurement_2.0-White_Paper.pdf

constructing a questionnaire that is brief yet psychometrically sound. It is beyond the scope of this article to describe the psychometric properties of all versions of the questionnaires. Therefore, we share here specific psychometric properties of the most commonly used ACORN questionnaire (see Table 1), as well as general psychometric indices that are obtained for all versions.[2] For more information, including examples of the used questionnaires, the reader is directed to the ACORN online questionnaire manual available at http://psychoutcomes.org/Questionnaires.

At the time of writing this manuscript (June, 2015), a total of 340,281 administrations of the Adult Version 11[3] were available, representing 115,882 episodes of care. The questionnaires were completed in a variety of outpatient treatment settings, including for-profit private practices, clinics affiliated with hospitals, and nonprofit agencies. Payer types include self-pay, commercial insurance, and Medicaid along with other sources of public funding. Adult Version 11 contains 13 items inquiring about some of the most common symptoms and problems reported by individuals seeking treatment. Ten of these items inquire about clinical symptoms, social isolations/conflict, and functioning in daily activities (i.e., Global Distress Scale [GDS]), while the remaining three are intended to be screening items for substance abuse. A factor analysis with varimax rotation conducted on the 13 items supported a two-factor solution, one consisting of the 10-item GDS factor and the other of the three remaining substance abuse screening items. Thus, pre-post effect sizes are calculated only using the GDS items. The correlation between GDS and substance abuse was $r = .17$, indicating that less than 3% of the variance are shared between the two scales. In addition, 4 items inquire about the therapeutic alliance, which will be discussed separately.

Included in Table 1 are the clinical cutoff score (CCS) and reliable change score (RCS) for the GDS, both based on Jacobson and Truax (1991). CCS was derived based on a normative sample collected during a pilot phase in 2007. For the GDS for the Adult Version 11, the CCS is based on 175 nonclinical participants who answered the particular 10 items (M [SD] = 1.13 [0.61]) and the 7,571 clients in treatment (M [SD] = 1.97 [0.79]) yielding a CCS of 1.50. Coincidentally, with this version as well as others, the CCS generally corresponded to the 25th percentile of the clients in treatment (i.e., 75% of those in treatment scored above the CCS).

RCS refers to the difference score between two administrations that is most likely (with one-tailed 95% confidence) attributable to change above and beyond measurement error, and is identical to the reliable change index (RCI) formula on p. 14 in Jacobson and Truax (1991) except that it is solved for the difference score:

$$x_2 - x_1 = \mathrm{RCI} \times S_{diff} = \mathrm{RCI} \times \sqrt{2\left(sd\sqrt{1 - r_{xx}}\right)^2}.$$

Here, sd and r_{xx} are, respectively, the standard deviation and reliability index of the GDS as reported in Table 1. As for the RCI, Jacobson and Truax uses a two-tailed 95% confidence (i.e., RCI = 1.960) although we believe that a one-tailed 95% confidence (equivalent to a two-tailed 90% confidence) is sufficiently reliable for our purposes (i.e., RCI = 1.645). RCSs calculated with both RCIs are reported in Table 1.

It is important to note that, upon consultation with the academic collaborators, the RCS is not directly applied to the raw pre-post effect size because it suffers from a statistical artifact in that whether or not a client shows "reliable change" depends heavily on the client's intake score simply as a function of regression to the mean (Hsu, 1989; Speer, 1992). Consequently, to classify clients as significantly improved, improved, unchanged, worse, or significantly worse, the RCS is applied to the severity adjusted effect size (SAES), which will be described in detail below. The intercept and slope from a regression analysis predicting the last GDS score based on the first GDS score are also included, which will be used for illustration purposes later when explaining SAES.

Data for concurrent validity have been collected through several participating organizations in the collaboration. With 3,903 concurrent administrations, the GDS and the PHQ9 (Kroenke & Spitzer, 2002) had a correlation of $r = .83$, indicating a 69% overlap in shared variance. Further, factor analysis with varimax rotation including all items of the GDS and the PHQ9 confirmed a single factor that accounted for 73% of the variance. The evidence for a single factor is consistent with other published research performing factor analyses on various commonly used outcome measures (Brophy, Norvell, & Kiluk, 1988; Lo Coco, Chiappelli, Bensi, Gullo, Prestano, & Lambert, 2008; Enns, Cox, Parker, & Guertin, 1998). We should note, however, that from the standpoint of clinical utility, the GDS is separated out into three subscales: clinical symptoms, social problems, and functioning/productivity in daily activities. Again, items asking about alcohol/drug use/abuse are scored as a separate scale.

Other versions have very similar psychometric properties due to the significant degree of item overlap of the most frequently used items. Questionnaires are constructed for the ACORN collaboration with a clear target reliability (Cronbach's alpha) of $\alpha = .90$ or higher which is typically achieved with 10–12 well-chosen items assessing anxiety, depression, sleep problems, concentration difficulty, social isolation/conflict, and social functioning. Factor analyses confirm that these items all load onto GDS. For adolescent and child measures, a few more items are typically needed to reach the same threshold of internal consistency.

All versions of the ACORN questionnaires also include items measuring therapeutic alliance. One psychometric caveat regarding the alliance items is that they are very highly skewed toward the clients reporting that their alliance is perfect. In other words, the default response for the client is to strongly agree with every positively worded alliance item. Thus, extreme caution is warranted when conducting typical parametric statistics using raw alliance scores (Minami, Wislocki, Brown, & Wampold, 2013). In the case of the questionnaire used for this article, 85% of the questionnaires rated all four alliance items as "perfect," with no room for improvement. The lack of a normal distribution, however, does not at all imply that these items are useless in either research or clinical practice—their impact on outcomes will be discussed in a subsequent section on findings from the ACORN collaboration. We also illustrate later how therapists could capitalize on the alliance items as opportunity to foster clients' honest feedback and enhance treatment outcomes.

[2] Psychometric information for specific versions is shared within the ACORN collaboration via the ACORN wiki web site.

[3] http://psychoutcomes.org/pub/Questionnaires/SampleQuestionnaires/AdultVer11.pdf

Table 1
Psychometrics of the Global Distress Scale (Adult Version 11)

Cases		
All cases		
N	Total administration	340,261
	Episodes of care	115,882
	Episodes with repeat administrations	55,921
Descriptive	*M* (*SD*)	2.07 (.78)
	Mean raw pre-post change	.40
	Mean raw pre-post effect size (Cohen's *d*)	.51
Reliability	Cronbach's alpha	.91
	Standard error of measurement	.23
	Clinical cutoff score	1.50
Clinical cases with repeat administrations		
N	Episodes of care	43,637
Descriptive	*M* (*SD*)	2.37 (.58)
	Mean raw pre-post change	.51
	Mean raw pre-post effect size (Cohen's *d*)	.89
Reliability	Standard error of measurement	.18
	Reliable change score[a] (reliable change index = 1.645)	.41
	Reliable change score (reliable change index = 1.960)	.49
Predictive	Intercept for last GDS score regressed onto first GDS score	.42
	Slope coefficient for last GDS score regressed onto first score	.61

Note. GDS = Global Distress Scale.
[a] Reliable change score (RCS) refers to the raw score differences between two administrations necessary to obtain a particular reliable change index (RCI; Jacobson & Truax, 1991). The ACORN collaboration uses a one-tailed 95% confidence interval (RCI = 1.645), whereas Jacobson and Truax reports a two-tailed 95% confidence interval (RCI = 1.960). RCSs are calculated with the respective RCIs.

SAES and Statistical Modeling

The ACORN work group, consisting of CCI members and academic researchers, had several methodological challenges to tackle in order to adopt Cohen's *d* as the reporting effect size unit. First, as noted earlier, simply due to regression to the mean, the magnitude of prepost change is directly proportional to the client's initial severity. Specifically, the correlation between the first and last observations is approximately $r = .50$ for all versions of the questionnaires, indicating that the initial severity accounts for roughly 25% of the prepost effect size. Therefore, it is crucial that any calculation of effect size as well as comparison among clients take into consideration their initial severity. Similarly, clients with intake scores below the CCS (i.e., 25th quartile or lower) averaged little to no improvement, with a significant percentage of clients showing a slight upward trend (i.e., seemingly getting worse). Further, given that we wanted to maintain comparability with results from clinical trials (i.e., benchmarking) and that most clinical trials use a cutoff score for participant inclusion, it was decided that effect sizes will only be calculated for clients with intake scores within the clinical range, that is, above the CCS. Second, raw effect sizes also do not adjust for additional clinical differences among clients, and in particular, diagnosis, although its magnitude is far less than initial severity (less than 1% of the variance in our data; consistent with Wampold & Imel, 2015). Therefore, rather than using the raw effect size as the reporting unit, we devised a SAES to account for these two primary clinical characteristics that affect treatment outcomes.

Specifically, SAES is calculated in two steps. First, a residual is calculated from the linear regression predicting the last GDS score based on the first GDS score and the client's diagnostic category. In other words, we calculate how the client fared at the last session controlling for their initial severity and diagnosis. Given the direction of the GDS (i.e., higher score indicates more symptoms), positive residual indicates that the client had symptoms that were worse than what could be expected based on severity and diagnosis, whereas a negative residual indicates that the client had less symptoms than expected. Second, to convert this relative indicator of well-being to pre-post effect size in Cohen's *d*, the residual is then subtracted (again, given the direction of the GDS) from the mean pre-post GDS change score and then divided by the standard deviation of the first GDS score.

For example, say Client A and Client B, both diagnosed with depression, attained a 0.50-point decrease from pre- to posttreatment. Given the standard deviation of the first GDS score of 0.58, they would both result in a raw effect size of $d = 0.86$ no matter what their initial severity was. However, if Client A's initial GDS score was 3.50 (and thus decreased to 3.00) and Client B's initial GDS score was 2.00 (and thus decreased to 1.50), their SAES would rightfully result in very different values. Specifically, based on the intercept and slope provided in Table 1, Client A's residual is $3.00 - (0.42 + 0.61 \cdot 3.50) = 0.45$, and thus SAES is $d = (0.51 - 0.45)/0.58 = 0.11$.. This is not a good outcome, in which most of the decrease in symptoms could be attributed to regression to the mean. On the other hand, Client B's residual is $1.50 - (0.42 + 0.61 \cdot 2) = -0.14$, and thus SAES is $d = [0.51 - (-0.14)]/0.58 = 1.12$. This is a very good outcome. By correctly accounting for differences in clinical conditions, the SAES reflects the reality that Client B made a significant improvement whereas Client A did not.

The importance of statistically adjusting the pre-post treatment effect size for clinical characteristics goes beyond comparing outcomes across individual clients and to comparing across therapists, clinics, and so forth This is because it is most likely that therapists, clinics, and so forth, vary significantly among one another with regards to their clients' clinical characteristics (i.e., case mix). For example, using the same numbers, say both Therapist A and Therapist B had the same number of clients and attained an average of 0.50-point decrease with their clients. If Therapist A had significantly more clients starting at the severe range, for example, on average an initial GDS score of 3.50, than Therapist B, whose average initial GDS score of their clients was 2.00, then it becomes clear that Therapist B on average has a much better treatment outcome than Therapist A. However, the two therapists' estimated average effect size will not be exactly $d = 0.11$ and $d = 1.12$, respectively, because of using a random effects model to estimate the therapists' effect sizes. If their averages were based only on a few clients each, their estimated effect sizes will be much closer to the overall mean (i.e., $d = 0.89$). On the other hand, if their averages were based on a significantly large number of clients (e.g., $n = 100$ each), their estimates will undoubtedly be very close to $d = 0.11$ and $d = 1.12$ (see Minami et al., 2012, for methodological details). Although the random effects model accounts for the uncertainty based on client sample size, so as to not potentially misinform the users, we do not provide estimates at the therapist's level unless there are at least 15 completed cases.

ACORN Toolkit Display

The results of the above calculations, and so forth, are provided to the clinicians, supervisors, managers, and so forth, in a user-friendly application. Figures 1 and 2 provide examples of the ACORN Toolkit screenshots that users have access via a secure website (www.psychoutcomes.org).

In Figure 1, the two histograms in the middle provide the SAES to the left and the raw effect size to the right. Although we provide the raw effect size based on our users' requests, we also explain that the raw effect size is problematic to interpret given the issues discussed earlier. For this clinician, his or her SAES is in the highly effective range (SAES ≥ 0.80; Minami et al., 2007, 2012), indicating that the clients under this clinician, on average, are doing very well. To the right of the histograms is a horizontal bar graph that breaks out the percentages of clients that the therapist saw that could be classified as significantly improved, somewhat improved, unchanged, somewhat worse, and worse, after having controlled for the case mix. The client cases are classified using (a) the standard error of measurement (SEM; GDS $= 0.18$; $d = 0.18/0.58 = 0.30$) plus a margin of $d = 0.20$ (defined by Cohen, 1988, as a "small" effect size) to ensure clinical effectiveness (thus $d = 0.30 + 0.20 = 0.50$; equals GDS $= 0.29$) and (b) the RCS (GDS $= 0.41$; $d = 0.41/0.58 = 0.70$). Clients who have had pre-post SAES change larger than the RCS are classified as significantly improved. Clients whose pre-post SAES change was

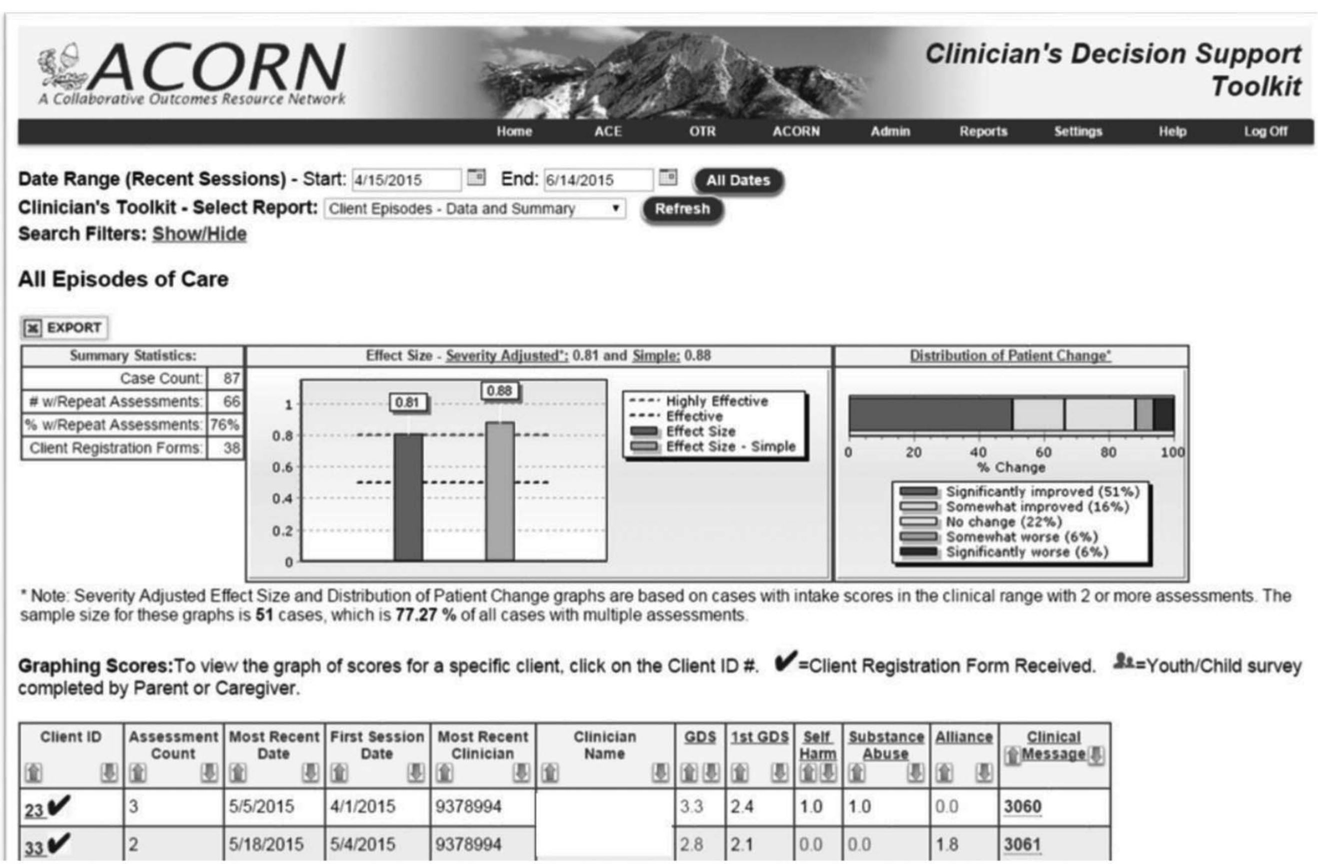

Figure 1. Screenshot of a clinician's view of the ACORN Toolkit.

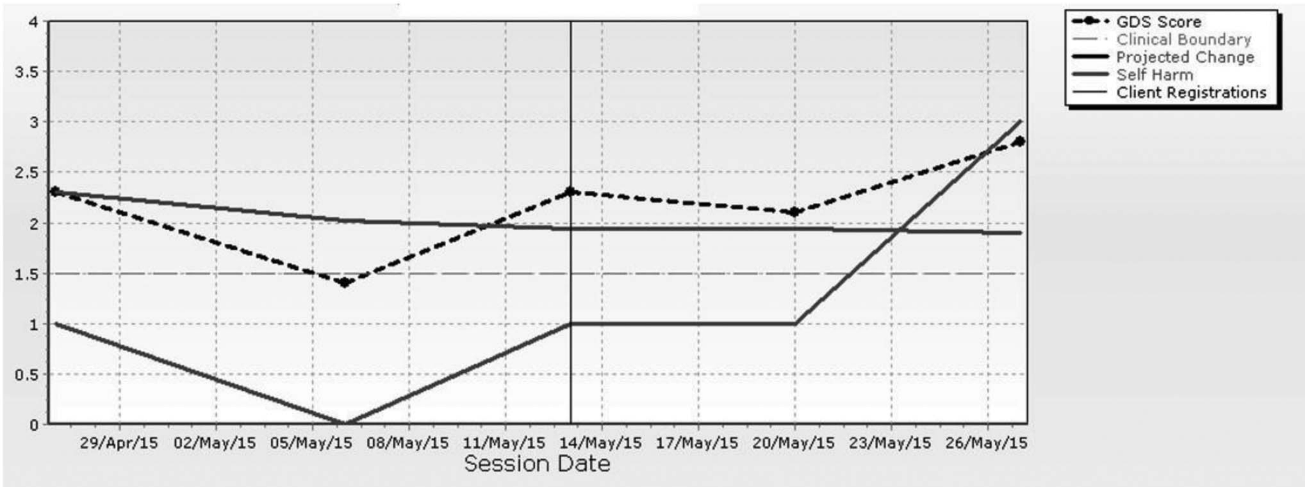

Figure 2. Graph provided in the ACORN Toolkit for an individual case with repeat assessments.

larger than the SEM plus the margin but smaller than the RCS are classified as somewhat improved. Conversely, clients with negative pre-post SAES change larger than the RCS and those between the RCS and the SEM plus the margin were respectively classified as significantly worse and somewhat worse. The remainder of the clients whose absolute magnitude of their pre-post SAES change is smaller than the SEM plus the margin is classified as unchanged.

Although we will not elaborate given space constraints, the ACORN Toolkit allows for the users (e.g., clinicians, supervisors, managers) to filter the clients' results by date range, age group, sex, diagnosis, and payer type. The Toolkit also enables the users to graph scores for individual cases by clicking on the client's ID (see Figure 2). In this instance, among numerous options, the graphed lines display the GDS score, self-harm ideation, and projected longitudinal trajectory based on a general linear model including intake score, session number, and weeks in treatment as predictors.

The ACORN Toolkit also provides an option to view the clinical message for the particular case based on the clinical algorithm that incorporates key clinical variables such as the GDS score, change since intake, number of sessions, self-harm ideation item response, change in self-harm ideation, substance abuse scale score, change in substance abuse scale score, alliance scale score, and changes in alliance score. Following is the message generated for the case example in Figure 2:

"This patient reports a relatively *severe level of distress* and *improvement is significantly less than average* in treatment compared to similar patients. *This pattern is associated with a risk of premature termination with a poor treatment outcome.*

The patient also reports a relatively high frequency of *suicidal ideation* that has increased since the start of treatment. This is associated with a *higher risk of hospitalization* and *premature termination* with a *poor treatment outcome."*

Findings From the ACORN Collaboration

The ACORN collaboration considers feedback informed treatment as the most literal form of evidence-based treatment—clients directly provide evidence to therapists with regards to whether or not their treatment is effective, and therapists, in return, have the opportunity to adjust their treatment accordingly. Consistent with this view, and as would be expected from research on measurement and feedback, ongoing analysis of ACORN data reveal a pattern of improving outcomes over time, in part mediated by the regularity with which therapists view their data.

Table 2 summarizes the results for therapists with at least 18 months of experience using the ACORN Toolkit. A total of 704 therapists are included, with SAES for 9,785 clients over the first 18 months of using ACORN questionnaires, and with another 30,410 treated in subsequent years after at least 18 months of experience. Both adults and children are included in the data, with 59% of the sample comprised of adults. Analysis of the relationship between frequency of Toolkit usage and improvement in outcomes indicated a nonlinear relationship in which therapists did not appear to benefit from Toolkit usage unless it was suggestive of routine use (i.e., 50 or more logins total over the 18-month period). For purposes of this article, the therapists were divided into low and high frequency usage based on whether or not they logged in more than 50 times.

The overall upward trend in effect size is evident for all users (moving from $d = 0.80$ to $d = 0.87$, a 9% gain), but particularly among high frequency Toolkit users (from $d = 0.82$ to $d = 0.95$, a 16% gain). Analysis of time spent online by high frequency users indicates that greater frequency of use resulted in even greater

Table 2
Therapists' Average Biannual Severity Adjusted Effect Size as Function of A Collaborative Outcome Resource Network Toolkit Login Frequency

Toolkit login (therapist *N*)	0–6 months (client *N*)	7–18 months (client *N*)	>18 months (client *N*)
Low (635)	.79 (1289)	.82 (6082)	.85 (22575)
High (69)	.82 (351)	.84 (2063)	.95 (7844)
Combined (704)	.80 (1640)	.82 (8145)	.87 (30419)

gains, where users who spent an average of approximately 10 min a week or more had the largest gains in SAES. Low frequency users also trended upward with an 8% gain in SAES, from $d = 0.79$ to $d = 0.85$. Multiple regression with number of months and Toolkit usage as predictors of SAES revealed that both months of experience and frequency of Toolkit usage are separate contributors to the observed gains in effect size ($p < .001$ for each predictor; $r = .029$ between the predictors, i.e., less than 0.1% variance overlap). During the initial 6 months, effect size differences between high and low frequency users are not statistically significant. However, for clients treated by these therapists after the initial 18 months of use of ACORN Toolkit, the difference in effect size between the high and low frequency users is statistically significant ($p < .001$). Of course, correlation is not causation; we cannot conclude that higher Toolkit usage will lead to better outcomes over time. Given that the therapists are not randomized, there might well be differences between therapists who readily use the Toolkit and those who do not. Further, as ACORN collaboration provides training and consultation through the use of live Webinars as well as online materials and videos, growing therapist comfort and skill with the questionnaires and alliance items may in part account for the upward trend in effect size observed even for those therapists who do not frequently log into the Toolkit to view results.

Consistent with the literature, there is a clear association between therapeutic alliance and treatment outcome. Although space limitations do not permit a full explanation of these findings, one notable finding is that clients who fail to complete alliance items tend to have significantly shorter lengths of treatment and smaller effect sizes. Table 3 presents outcomes for the same sample of clients based simply on whether they completed the alliance items at the final session or left them blank. Observed differences between those who complete alliance items and those who do not are evident for SAES, questionnaire count, and duration of treatment and are all statistically significant ($p < .001$).

Another finding that has been consistent is that clients who report that the alliance is perfect at every session do not necessarily have the best outcomes. The clients who report perfect or near perfect alliance at every session with no change from beginning to end of treatment had a mean SAES of $d = 0.77$. The pattern associated with the largest effect size is one in which the client reports less than perfect alliance early in treatment and then reports improvement in alliance by the end of therapy ($d = 0.86$). Approximately 25% of clients fit this pattern. Conversely, with approximately 20% of clients who exhibit a pattern of reporting perfect or near perfect alliance at the start of treatment but lower alliance by the end, the effect size is substantially lower ($d = 0.57$).

Table 3
Severity Adjusted Effect Size, Questionnaire Count, and Treatment Duration by Completion of Alliance Items at Last Session

Alliance items completed	*N*	SAES	Questionnaires (*N*)	Duration (weeks)
Yes	35965	.87	6.5	17.5
No	4239	.65	4.8	15.3

Note. SAES = severity adjusted effect size.

Implications

The above findings strongly suggest that ACORN Toolkit usage and skillful use of the alliance feedback is associated with larger effect sizes. First, it is important that the client feel comfortable with providing the feedback. To this end, the therapist is encouraged to explain at the very beginning of treatment that the ongoing feedback on both symptom improvement and alliance is helpful in achieving good outcomes. In particular, therapists should encourage their clients to be honest with them regarding their performance—feedback that the alliance is less than perfect should especially be greeted with openness and appreciation on the part of the therapists, thus communicating respect for the client's point of view and willingness to take the feedback seriously. Below, we provide specific suggestions based on different user role and context.

Use in Clinical Practice and Training

In clinical practice and training, the above findings clearly have implications on client progress monitoring and clinical supervision. First, therapists are encouraged to use the ACORN Toolkit to monitor the results of their clients. Second, clinical supervisors and managers are encouraged to do the same for their treatment team using the Toolkit to identify cases that suggest further review and supervision. The ACORN Toolkit is used in a wide variety of clinical settings, including mental health agencies, group practices, hospitals, and clinics associated with graduate training programs in psychology and related fields. Although some users are solo practitioners, the vast majority is part of a larger practice, and thus the Toolkit provides ability for supervisors and managers to see aggregate results of their therapists.

Individual therapists are able to closely monitor their individual cases to quickly identify active clients who are at highest risk for a poor outcome. In such cases, therapists are encouraged to discuss the poor treatment response with their clients, such as by sharing, "It seems that therapy has not been working very well for you. Are there things that I could do differently that may help you feel better?" Or, it may be that the therapist is unaware of the full extent of the client's issue and could ask: "Is there something else that may be going on in your life?" In these cases, it is of course crucial that the clients do not feel blamed for the lack of treatment response. In addition, alliance items may play a key role, especially if there is a decrease in rating or the client has opted not to complete the alliance items. In these situations, the therapist could comment: "It appears that I haven't been understanding you very well. Could you share with me what I might be missing?" In addition, individual therapists also have the ability to view their aggregated results and compare their SAES to norms developed by the ACORN collaboration. This allows therapists to compare how their clients are doing on average as compared to the SAES norms as well as how they, as therapists, are doing compared to other therapists. Especially for therapists who are doing very well, it would clearly be to their advantage to show their results to their employers or to third-party payers so as to negotiate their reimbursement rate with solid evidence of performance.

Clinical directors and supervisors are provided access to data for all therapists on their team, and can likewise look closely and identify at-risk cases. These cases can be then bought up as part of supervision or team meetings; in fact, it would allow supervisors to prioritize which cases to discuss rather than relying on the supervisees to decide which cases to share. For example, "It appears that the client's suicidal ideation increased this past week. What has been going on with the client, and how have you been working with him/her?" In such cases, the supervisor would want to review the case with the therapist in depth, with particular attention to what steps are need to address the increase in self harm ideation and maximize the chances that the client remains in treatment. With cases that their supervisees seem to be struggling with, supervisors could actively explore both clinical symptoms and alliance in association with how the supervisee has been working with the client: "It seems like this client has not improved much over the past 2 months. How have you been working with this client? What might we need to do differently?" By knowing which clients are of priority, supervision could devote more time to clients who are in need.

Practice owners and agency directors are also able to view results for the entire practice/agency, and view how their results compare to other similar sites using the ACORN Toolkit. Not only could this information be used for marketing purposes to promote more business, agencies can also leverage this knowledge to negotiate contracts with third-party payers or secure grants from nonprofit organizations and other funding sources. Some sites have gone so far as to include therapist and supervisor usage of the Toolkit as quality indicators for their quality improvement initiatives, as well as provide incentives for increased Toolkit usage.

Implications for Future Training

Recent studies in psychotherapy have clearly demonstrated that significantly larger amount of variance in treatment outcomes is due to the therapist than to the method of therapy (Wampold & Imel, 2015). In concert with the evidence that therapists are not the best judge of their own outcomes (Dunning, Heath, & Suls, 2004; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Hannan et al., 2004; Walfish, McAlister, O'Donnell, & Lambert, 2012), it logically suggests that if we wish to improve treatment outcomes, we need to provide decision support tools to therapists and therapists in training. Feedback informed treatment clearly achieves this aim. Routine use of outcome and alliance questionnaires combined with continuous feedback to therapists results in improved outcomes, and is evidenced by the following literature reviews and meta analyses (Lambert, 2010a; Lambert, 2010b; Shimokawa, Lambert, & Smart, 2010; Goodman, McKay, & DePhilippis, 2013). However, the wide variability in effect sizes across therapists raises new challenges for psychotherapy training. Simply put, how effective does a new therapist need to be on average in order to be considered qualified to independently treat clients? Does a supervisor need to demonstrate equivalent or better outcomes than the supervisees? Should a training program be judged by its ability to produce graduates with strong practice-based evidence of effectiveness? Whether or not therapists should measure outcomes of their clients is no longer a question, but is an ethical issue—if not, on what basis could any therapist argue that s/he is very effective at what s/he does? As the literature suggests, what the therapists claim they are doing in treatment (e.g., theoretical orientation) provides no substantial evidence (e.g., Brosan, Reynolds, & Moore, 2008).

This is also true with training—if a training program is not measuring and evaluating their trainee's outcomes, on what basis could the training program argue that they are teaching evidence-based treatment? Two recent studies looking at outcomes for clinics serving as training sites for graduate students found significant differences between therapists, but no evidence that the experienced and licensed clinical staff had better outcomes than the graduate trainees (Minami et al., 2009; Okiishi, Lambert, Nielsen, & Ogles, 2003). The belief that graduate training and traditional supervision alone will produce effective therapists is not supported by the evidence. Instead, the evidence supports the conclusion that whatever factors do contribute to the therapist's effectiveness are not being enhanced by traditional training. Failure to expose graduate students to research on therapist effects, evidence for feedback informed treatment, and routine clinical training with questionnaires along with performance feedback breeds the next generation of therapists who believe, yet lacks any substantial evidence, that they are effective.

Furthermore, based on the evidence from various clinical trials on the importance of feedback and measurement in achieving superior outcomes, as well as the findings from the ACORN collaboration reported herein, it is reasonable to deduce that graduates trainees who are not exposed to measurement and feedback may not be as effective as they could have been if they had been trained using feedback informed treatment. Whipple et al. (2003) found that clients whose therapists had access to progress and alliance information were less likely to deteriorate, more likely to stay longer, and twice as likely to achieve a clinically significant change. Further, Duncan, Miller, Wampold, and Hubble (2010) expand upon this research in their introduction to their book-length comprehensive review of evidence based practices in mental health care.

Thus, we believe that the evidence requires the introduction of feedback informed treatment into all therapist training programs if they wish to claim that they are training effective therapists based on evidence. Without this experience and comfort with ongoing performance feedback, the new therapists will not be prepared to thrive in a world that demands accountability and evidence of effectiveness. Without ongoing performance feedback, therapists lack the information necessary to improve their results—or worse, they may simply remain ineffective or even deteriorate. Consequently, organizations that accredit training programs should require evidence-based practice in the form of feedback informed treatment as part of the accreditation standards. At the time of this writing, two academic programs have decided to use the ACORN Toolkit for clinical training. We hope that eventually all academic programs would incorporate true evidence based practice into their training for the sake of their students and their clients.

References

Brophy, C. J., Norvell, N. K., & Kiluk, D. J. (1988). An examination of the factor structure and convergent and discriminant validity of the SCL-90R in an outpatient clinic population. *Journal of Personality Assessment, 52,* 334–340. http://dx.doi.org/10.1207/s15327752jpa5202_14

Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy, 36,* 581–587. http://dx.doi.org/10.1017/S1352465808004438

Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J. (2001). Pushing the quality envelope: A new outcomes management system. *Psychiatric Services, 52,* 925–934. http://dx.doi.org/10.1176/appi.ps.52.7.925

Brown, G. S., & Lambert, M. J. (1998, June). *Tracking patient progress: Decision making for cases who are not benefiting from psychotherapy*. Paper presented at the 29th Annual Meeting of the Society for Psychotherapy Research, Snowbird, UT.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Duncan, B., Miller, S., Wampold, B., & Hubble, M. (2010). Introduction. In B. Duncan, S. Miller, B. Wampold, & M. Hubble (Eds.), *The heart and soul of change: Delivering what works in therapy* (2nd ed., pp. 23–42). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12075-000

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69–106. http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87. http://dx.doi.org/10.1111/1467-8721.01235

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105,* 98–121. http://dx.doi.org/10.1016/j.obhdp.2007.05.002

Enns, M. W., Cox, B. J., Parker, J. D., & Guertin, J. E. (1998). Confirmatory factor analysis of the Beck Anxiety and Depression Inventories in patients with major depression. *Journal of Affective Disorders, 47,* 195–200. http://dx.doi.org/10.1016/S0165-0327(97)00103-1

Goodman, J. D., McKay, J. R., & Dephilippis, D. (2013). Progress monitoring in mental health and addiction treatment: A means of improving care. *Professional Psychology: Research and Practice, 44,* 231–246. http://dx.doi.org/10.1037/a0032605

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61,* 155–163. http://dx.doi.org/10.1002/jclp.20108

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment, 11,* 459–467.

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. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32,* 1–7. http://dx.doi.org/10.3928/0048-5713-20020901-06

Lambert, M. J. (2010a). *Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12141-000

Lambert, M. J. (2010b). Yes, it is time for clinicians to routinely monitor treatment outcome. In B. Duncan, S. Miller, B. Wampold, & M. Hubble (Eds.), *The heart and soul of change: Delivering what works in therapy* (2nd ed., pp. 239–266). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12075-008

Lambert, M. J., & Brown, G. S. (1996). Data-based management for tracking outcome in a private practice. *Clinical Psychology: Science and Practice, 3,* 172–178. http://dx.doi.org/10.1111/j.1468-2850.1996.tb00068.x

Lo Coco, G., Chiappelli, M., Bensi, L., Gullo, S., Prestano, C., & Lambert, M. J. (2008). The factorial structure of the Outcome Questionnaire-45: A study with an Italian sample. *Clinical Psychology & Psychotherapy, 15,* 418–423. http://dx.doi.org/10.1002/cpp.601

Minami, T., Brown, G. S., McCulloch, J., & Bolstrom, B. J. (2012). Benchmarking therapists: Furthering the benchmarking method in its application to clinical practice. *Quality & Quantity: International Journal of Methodology, 46,* 699–708. http://dx.doi.org/10.1007/s11135-011-9548-4

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. http://dx.doi.org/10.1037/a0015398

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

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S., & Kircher, J. C. (2008b). 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. http://dx.doi.org/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. http://dx.doi.org/10.1037/0022-006X.75.2.232

Minami, T., Wislocki, A. P., Brown, G. S., & Wampold, B. E. (2013, July). Modeling psychotherapy treatment progress and therapeutic alliance in natural clinical settings. In Z. E. Imel (Chair), *Interpersonal influence processes in psychotherapy*. Symposium presented at the 121st Annual Convention of the American Psychological Association, Honolulu, HI.

Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work II. *Psychotherapy, 48,* 4–8. http://dx.doi.org/10.1037/a0022180

Okiishi, J., 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. http://dx.doi.org/10.1002/cpp.383

Shimokawa, K., Lambert, M. J., & Smart, D. W. (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. http://dx.doi.org/10.1037/a0019247

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991). revisited. *Journal of Consulting and Clinical Psychology, 60,* 402–408. http://dx.doi.org/10.1037/0022-006X.60.3.402

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports, 110,* 639–644. http://dx.doi.org/10.2466/02.07.17.PR0.110.2.639-644

Wampold, B. E., & Brown, G. S. J. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73,* 914–923. http://dx.doi.org/10.1037/0022-006X.73.5.914

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). New York, NY: Routledge.

Wells, M. G., Burlingame, G. M., Lambert, M. J., Hoag, M., & Hope, C. (1996). Conceptualization and measurement of patient change during psychotherapy: Development of the outcome questionnaire and youth outcome questionnaire. *Psychotherapy: Theory, Research, Practice, Training, 33,* 275–283. http://dx.doi.org/10.1037/0033-3204.33.2.275

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 failure and problem solving strategies in routine practice. *Journal of Counseling Psychology, 50,* 59–68. http://dx.doi.org/10.1037/0022-0167.50.1.59

Received July 1, 2015
Accepted July 2, 2015 ■

New Editors Appointed

The Publications and Communications Board of the American Psychological Association announces the appointment of 6 new editors. As of January 1, 2016, manuscripts should be directed as follows:

- *American Psychologist* (www.apa.org/pubs/journals/amp/) **Anne E. Kazak, PhD, ABPP,** Nemours Children's Health Network, A.I. du Pont Hospital for Children
- *Developmental Psychology* (http://www.apa.org/pubs/journals/dev/) **Eric F. Dubow, PhD,** Bowling Green State University
- *International Perspectives in Psychology: Research Practice, Consultation* (www.apa.org/pubs/journals/ipp/) **Stuart Carr, PhD,** Massey University
- *Journal of Consulting and Clinical Psychology* (www.apa.org/pubs/journals/ccp/) **Joanne Davila, PhD,** Stony Brook University
- *School Psychology Quarterly* (www.apa.org/pubs/journals/spq/) **Richard Gilman, PhD,** Cincinnati Children's Hospital Medical Center
- *Sport, Exercise and Performance Psychology* (www.apa.org/pubs/journals/spy/) **Maria Kavussanu, PhD,** University of Birmingham, UK

Electronic manuscript submission: As of January 1, 2016, manuscripts should be submitted electronically to the new editors via the journal's Manuscript Submission Portal (see the website listed above with each journal title).

Current editors Norman Anderson, PhD, Jacquelynne Eccles, PhD, Judith Gibbons, PhD, Arthur M. Nezu, PhD, Shane R. Jimerson, PhD, and Jeffrey J. Martin, PhD will receive and consider new manuscripts through December 31, 2015.