Psychotherapy Research

ISSN: 1050-3307 (Print) 1468-4381 (Online) Journal homepage: http://www.tandfonline.com/loi/tpsr20

Effectiveness of routine psychotherapy: Method matters

Andrew A. McAleavey, Soo Jeong Youn, Henry Xiao, Louis G. Castonguay, Jeffrey A. Hayes & Benjamin D. Locke

To cite this article: Andrew A. McAleavey, Soo Jeong Youn, Henry Xiao, Louis G. Castonguay, Jeffrey A. Hayes & Benjamin D. Locke (2017): Effectiveness of routine psychotherapy: Method matters, Psychotherapy Research, DOI: 10.1080/10503307.2017.1395921

To link to this article: http://dx.doi.org/10.1080/10503307.2017.1395921

Published online: 02 Nov 2017.

Psychotherapy Research, 2017
https://doi.org/10.1080/10503307.2017.1395921

EMPIRICAL PAPER

Effectiveness of routine psychotherapy: Method matters

ANDREW A. MCALEAVEY [1], SOO JEONG YOUN[2], HENRY XIAO [2], LOUIS G. CASTONGUAY[2], JEFFREY A. HAYES [3], & BENJAMIN D. LOCKE[4]

[1]*Department of Psychiatry, Weill Cornell Medical College, New York, NY, USA;* [2]*Department of Psychology, The Pennsylvania State University, State College, PA, USA;* [3]*Department of Counseling Psychology, The Pennsylvania State University, State College, PA, USA &* [4]*Counseling and Psychological Services, The Pennsylvania State University, State College, PA, USA*

(Received 16 September 2016; revised 12 October 2017; accepted 13 October 2017)

ABSTRACT

Objective: Though many studies have shown that psychotherapy can be effective, psychotherapy available in routine practice may not be adequate. Several methods have been proposed to evaluate routine psychological treatments. The aim of this paper is to demonstrate the combined utility of complementary methods, change-based benchmarking, and end-state normative comparisons, across a range of self-reported psychological symptoms. **Method:** Benchmarks derived from randomized controlled trials (RCTs) and normative comparisons were used to evaluate the effectiveness of psychotherapy in a large (N = 9895) sample of clients in university counseling centers (UCCs). **Results:** Overall, routine psychotherapy was associated with significant improvement across all symptoms examined. For clients whose initial severity was similar to RCT participants, the observed pre–post effect sizes were equivalent to those in RCTs. However, treatment tended to lead to normative end-state functioning only for those clients who were moderately, but not severely, distressed at the start of psychotherapy. **Conclusions:** This suggests that although psychotherapy is associated with an effective magnitude of symptom improvement in routine practice, additional services for highly distressed individuals may be necessary. The methods described here comprise a comprehensive analysis of the quality of routine care, and we recommend using both methods in concert.

Keywords: psychotherapy; psychotherapy effectiveness; counseling; benchmarking; normative comparisons

Clinical or methodological significance of this article: This study examines the effectiveness of routine psychotherapy provided in a large network of counseling centers. By comparing multiple established methods to define outcomes in this sample we provide a detailed understanding of typical outcomes. The findings show that, across several different problem areas, routine psychotherapy provided substantial benefit, particularly to clients in the most distress. However, there is room to improve, especially by increasing the number of clients who return to normal functioning by the end of treatment. Using distinct methods provides complementary answers to the question: How effective is routine psychotherapy?

Several meta-analyses of psychotherapy for various psychiatric disorders, such as major depressive disorder (MDD), obsessive-compulsive disorder, posttraumatic stress disorder and many others (see e.g., Cuijpers et al., 2013, 2014) have found that some psychotherapies are more effective than placebo controls. From these findings, however, one cannot conclude whether psychotherapy delivered in routine care is as effective as therapies examined in controlled settings—or even helpful at all. Several authors have raised concerns about the quality of psychotherapy available to most clients outside of research trials, noting that a non-trivial portion of clients actually deteriorate during therapy (Lilienfeld, 2007). However, standard randomized controlled trials (RCTs), which provide strong evidence in specific cases, may not represent the reality of the treatments applied in the real world. Aside from deviations from evidence-based interventions which

Correspondence concerning this article should be addressed to Andrew A. McAleavey, Department of Psychiatry, Weill Cornell Medical College, New York, NY 10065, USA. Email: andrew.mcaleavey@gmail.com

might occur in practice, these research methods are limited in external validity and provide strong inferences only regarding specific treatments, settings, and populations. To assess the quality of psychotherapy outcomes in routine care requires distinct methods to account for the many differences between routine and clinical research environments, including: lack of random assignment, more greatly varied treatment types offered, and less well-defined inclusion/exclusion criteria.

One important development in defining "effectiveness" across treatments is Jacobson and Truax's (1991) clinically significant change (CSC), which was proposed as a common assessment of the effects of treatment on clients' functioning and a way to determine whether an individual patient benefitted from treatment. However, there are several factors limiting the applicability of Jacobson and Truax's methods directly to the study of routine care. First, their original and widely accepted definition of clinical significance is at the individual level, and the calculation required to aggregate multiple individuals frequently becomes complex because multiple categories of outcomes may be considered (e.g., Ronk, Korman, Hooke, & Page, 2013). Furthermore, a number of formulae are available to calculate the RCI and cut points, making their application variable across studies. The RCI, in particular, is highly sensitive to various methodological choices (including which reliability estimates to use) and sample-specific attributes (such as initial severity and pre-treatment standard deviation; e.g., Barkham, Stiles, Connell, & Mellor-Clark, 2012). Thus, although CSC may be helpful in documenting effectiveness, the Jacobson and Truax methods are limited in determining whether routine treatments are effective because they lack robust comparisons.

In recent years, two other methods have been used to assess the effectiveness of treatment outcomes outside of RCTs: benchmarking routine psychological treatment outcomes and normative comparisons. These two methods can be conceived to represent a group-level adaptation of Jacobson and Truax's (1991) initial two-part definition of CSC: a change component and an end-state component, and they have the benefit of including formal tests for equivalence.

Benchmarking Routine Psychological Treatment Outcomes

The overall strategy of benchmarking is to compare aggregate effect sizes (or other metrics) derived from routine care to carefully selected studies (the benchmarks), in order to determine if the observations made in routine care are roughly equivalent to what might be expected in RCTs. Though this comparison is inherently limited due to the very different contexts of a controlled trial and routine care, increasingly formal methods have been developed to perform more valid tests.

The first studies to informally include benchmarking (e.g., Forand, Evans, Haglin, & Fishman, 2011; Hunsley & Lee, 2007; Merrill, Tolbert, & Wade, 2003; Wade, Treat, & Stuart, 1998) lacked statistical comparisons between routine care and benchmark studies: they provided the effect sizes for each, and because the two are close to each other (and in some cases the routine care actually yielded a larger effect size), the authors conclude that treatments were generally effective. Weersing and Weisz (2002)'s formal approach, which provided a statistical framework using meta-analytic methods to compare naturalistic and RCT samples, has two advantages over less formal approaches. First, it can accommodate many comparison studies through meta-analytic aggregation. Second, with formal statistical tests, levels of confidence can be calculated for any given comparison, which allows for greater objectivity.

Minami, Serlin, Wampold, Kircher, and Brown (2008), Minami, Wampold, Serlin, Kircher, and Brown (2007) provided a more general and formal method of developing benchmarks for routine practice, and applied that method to studies of adult MDD. Their method includes at least three major advances beyond other methods of benchmarking. First, they identified separate benchmarks for different categories of outcome measures: they found that high reactive measures (therapist-rated) showed larger effect sizes than low reactive measures, and that high specificity (symptom-specific) measures showed larger effect sizes than low specificity (general mental health or functioning) measures. This addresses a possible limitation in Weersing and Weisz's (2002) study due to the aggregation of scores across outcome type, which confounds method variance and outcome variance.

A second advantage of the Minami and colleagues' (2007, Minami, Serlin, et al., 2008) method is that, by using a range-null hypothesis test, they developed a formal statistical test for "clinical equivalence" to a research benchmark. Specifically, they set the benchmark target so as to allow an inference on whether an observed naturalistic sample was within 0.2 (on the effect size *d* scale) of the target efficacy benchmark, which would indicate that the naturalistic sample is not meaningfully different than the efficacy benchmark. This maintains the clinical rather than statistical idea of equivalence as primary, consistent with earlier notions of CSC: routine outcomes should be neither punished nor lauded for producing

statistically equivalent results to RCTs. Instead, routine outcomes should rather be evaluated based on whether the clinical outcomes of the treatments are equivalent to RCTs.

Finally, the Minami et al. (2007) method also incorporates a natural recovery benchmark, developed by creating a benchmark based on the waitlist control condition from the same RCTs. Using the same range-null hypotheses tests, they suggest that change in routine treatment should be greater than 0.2 *d* higher than the waitlist benchmark, and that if this is true, the routine treatment can be considered more effective than no-treatment.

Minami, Wampold, et al.'s (2008) benchmarking study compared a low-reactive, low-specificity measure (the Outcome Questionnaire-30 (OQ-30); Lambert et al., 2003) to the efficacy benchmarks established for bona fide treatments of MDD. Specifically, these authors used a large ($N = 5704$) data set collected from a managed care setting and compared various subsets of the data to clinical trials efficacy. The results showed that, for individuals who started treatment above the clinical cut score on the OQ-30, the routine treatments were clinically superior to no-treatment and clinically equivalent to efficacy trial treatments. Minami et al. (2009) used similar methods to evaluate a single counseling center's treatment efficacy (TE), and found that, for clients who started treatment elevated in distress, the amount of change observed in this center was clinically equivalent to efficacy benchmarks for MDD and clinically superior to natural history (NH).

The studies of benchmarking presented here, especially the Minami et al. (Minami, Wampold, et al., 2008; Minami et al., 2009) studies, raise an important question for the field of psychological treatments. Although the lack of randomization in these studies prevents inference about whether psychotherapy is the causative factor, these studies certainly suggest that, on average, routine treatments can be as effective as what are usually deemed the "state-of-the-art" (i.e., scientifically supported) treatments, performed by expertly trained and supervised therapists in RCTs. If this is true—and there are valid scientific reasons to be skeptical of such claims—it may mean that further dissemination of evidence-based interventions might only produce small to negligible gains in therapeutic outcomes, because psychotherapists are already as effective as RCT protocol therapists.

However, there are several possible limitations to be considered. For instance, it could be argued that even the most formal method of benchmarking change (Minami et al., 2007) is not a comprehensive method. After all, the amount of change observed is only one portion of the relevant information about the outcome of treatments. The symptom level of individuals at post-treatment (end-state functioning) is an arguably meaningful point of comparison. Another important limitation is the difficulty determining whether a treatment can be described as the cause of improvement, or if the change can be attributed to outside influences or statistical artifacts. For example, one important statistical artifact is regression to the mean, which in this case would describe the commonly observed phenomenon that larger effect sizes are generally observed for individuals who enter treatment at higher levels of symptoms. This can be ameliorated by a waitlist or NH control condition, provided random assignment is used, but the use of a cut score to separate clinical from nonclinical participants (as is common in these studies; e.g., Barkham et al., 2012; Minami et al., 2009) creates an inflated initial mean symptom score at pre-treatment, and is thus subject to regression to the population average over time.[1]

Another limitation is that, even though average changes in general psychological symptoms suggest equivalence between clinical routine and RCTs in the Minami et al. (2009; Minami, Wampold, et al., 2008) studies, specific symptoms that are targeted in treatment (e.g., symptoms of depression or eating disorders) may not show the same pattern. As described by McAleavey, Nordberg, Kraus, and Castonguay (2012), specific measures cannot be inferred from general ones. Clinically, overall or general distress may adequately capture many clients' reported concerns, but for some target problems, a general distress factor may not be directly related to success or failure in therapy (for example, when a client reports low distress except for in a specific area, like disordered eating). For these reasons, it is necessary to consider multiple types of client problems in treatment, and adapt the Minami et al. (2007) method to separate benchmarks for distinct problem types.

These limitations were partially addressed by McEvoy and Nathan (2007), who compared 143 outpatients in a mixed-diagnosis group CBT in routine care to efficacy trial groups of depression and anxiety patients. One important novel method in this paper was to contrast results based on pre–post change effect size benchmark analyses with results based on reliable change and CSC, the latter incorporating an end-state functioning comparison. The authors reported that their sample's effect size was close to the effect size of clinical trials for anxiety, depression, and mixed samples. However, their analyses showed a relatively small number of patients meeting Jacobson and Truax's (1991) definition of "recovered" on the measure of anxiety as compared

to depression, perhaps indicating the importance of assessing separate constructs rather than averaging across symptom clusters. Importantly, there was no formal test of whether the routine treatments produced lower percentages of recovery than the benchmark studies.

As pointed out by McEvoy and Nathan (2007), studies of clinical benchmarking do not address end-state functioning, which is an important consideration in evaluating routine practice. If routine treatments are producing equivalent change to RCT trials but inadequate end-state results, it is possible to conclude that benchmarking methods are not an adequate measure of clinical effectiveness.

Normative Comparisons

Another method to assess the effectiveness of treatment outcomes outside of RCTs has been proposed by Kendall and colleagues (Kendall & Grove, 1988; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999; Sheldrick, Kendall, & Heimberg, 2001). Their method of normative comparisons can be used as a formal statistical test of an important end-state question: Are clients in routine care within the range of "normal" or "healthy" persons at the end of treatment?

Any method of defining "normal" with regard to psychological symptoms will likely be controversial to some degree. However, Kendall and Grove (1988) suggested a reasonable and common guideline for this purpose. In order to account for the fact that even people who are neither in treatment nor seeking help may experience significant symptoms of psychopathology, they suggested that the mean of a normative sample ±1 SD be considered the "normal" range of functioning. The specific target of 1 SD from the mean was selected because it is common in clinical instruments to use this as the range considered "average." This comparison does not test for elimination of symptoms for the clinical group, but whether the treatment has brought the participants' average level of distress into a range more typical of the general population. Using statistical equivalence testing, Kendall and colleagues (1999) provided a test for whether a group average post-treatment score is reliably within this normative range. Because Kendall's normative comparisons were developed specifically for analysis of group data, the method may be advantageous for the analysis of routine care. Additionally, as noted above, while results of the Jacobson and Truax method are highly susceptible to choices made by researchers, Kendall et al.'s normative comparison method is more dependent on a separate normative sample, and thus should be a more fixed and stable criterion across studies.

Several improvements to the normative comparison methods suggested by Kendall et al. have been made by Cribbie and colleagues (Cribbie & Arpin-Cribbie, 2009; Mangardich & Cribbie, 2014; Nasiakos, Cribbie, & Arpin-Cribbie, 2010; van Wieringen & Cribbie, 2014). Specifically, Cribbie and Arpin-Cribbie accommodated unequal variances across groups and proposed a new hierarchical procedure for testing normative comparisons. Mangardich and Cribbie (2014) proposed a method of normative comparison that is robust not only to heteroskedasticity, but also to non-normally distributed data (which is often the case in post-treatment samples, and clinical instruments as well). This method (referred to as the Schuirmann-Yuen, or S-Y, method) calculates trimmed means and Winsorized variances, which exclude extreme values from strongly non-normal distributions. van Wieringen and Cribbie (2014) conducted a Monte Carlo simulation study comparing the methods proposed by Kendall against the S-Y method. The results showed that, in almost all circumstances (including unequal sample sizes, different distribution shapes, and different variances) the methods were either equivalent or showed distinct advantages for the S-Y method in Type I and Type II error control; the other methods were often intolerably error-prone. Based on this, the authors recommended the use of the S-Y method.

The Present Study

The goal of this paper is to compare and demonstrate the combined utility of change-based benchmarking and end-state normative comparison methods to assess routine psychological treatment. These methods offer different and complementary meanings of effectiveness, and using both methods in tandem allows us to address two distinct but related questions: (1) Is the amount of change produced in routine care equivalent to that seen in benchmarked trials?; and (2) Are clients' distribution of post-treatment scores statistically non-distinguishable from a normative sample? If a routine treatment affirms both of these questions, a very strong argument can be made that the treatment is as effective as can reasonably be expected, given current best practices. Should one or both questions return a negative finding, an argument may be made that improvements to regular practice may be possible, for instance through additional training in specific techniques and treatments, additional sessions provided

to clients in need of further treatment, or other administrative changes.

In addition, this paper addresses the question of whether there are specific clinical problems that might be more or less adequately treated in routine practice. Many studies of effectiveness are targeted to specific clinical diagnoses (e.g., Merrill et al., 2003; Minami, Wampold, et al., 2008). While this strategy provides excellent information about subpopulations of clients meeting certain criteria, it fails to examine the effect of routine practice on the general population of clients in therapy. Alternatively, many of the studies that have examined clients in routine care (e.g., Barkham et al., 2012; Minami et al., 2009) have tracked improvement using general symptom distress measures, which cover many domains of psychological functioning but may not have a direct clinical interpretation other than "distress" or "symptoms." Thus, in this paper, we assessed change across several symptom domains using a multidimensional measure of psychological symptoms, rather than focus on a single dimension of functioning. In sum, we not only compare two methods of operationalizing clinical effectiveness but also test whether certain symptoms are more effectively treated than others under the conditions of treatment as usual (TAU), with the ultimate aim of providing a broad evaluation of the effectiveness of routine treatment.

Method

Participants

Participants were clients seeking treatment at university counseling centers (UCCs) that were members of a large practice-research network, the Center for Collegiate Mental Health (CCMH). This network is described in detail in McAleavey, Lockard, Castonguay, Hayes, and Locke (2015) . Briefly, at participating centers, standardized data from each client including demographics, treatment history, and symptomatic outcomes is collected and aggregated directly through electronic medical records (EMR) software. Appropriate consent is collected at each UCC according to an institution-specific IRB approval. CCMH data are separated into academic-year data sets; for this study, the 2010–2011 and 2011–2012 data sets were combined in order to increase sample size.

Data Reduction

The initial data set included 161,335 clients, seen by 3,359 therapists, at 122 different UCCs. However, the data set includes heterogeneity in terms of services offered and frequency of data collected based on each center's policy (i.e., some centers may only collect data at intake, others more often). Thus, this large initial data set needed to be reduced. The data reduction steps are shown in Figure 1. Inclusion and exclusion rules were set a priori to identify those clients who were likely enrolled in individual, face-to-face psychotherapeutic treatments, rather than other forms of services (e.g., group formats, exclusively pharmacological treatments, academic counseling, or exclusively assessment services), and for whom adequate symptom data were collected. This is necessarily a probabilistic judgement, since aggregate EMR data of this type is imperfect.

We first reviewed the appointment descriptors for each appointment in the data set and determined whether that appointment was potentially part of a routine individual face-to-face psychological treatment, or whether it was better considered a separate service or supplemental treatment. Three authors (AUTHOR INITIALS REMOVED FOR BLIND REVIEW) independently coded all 1,726 unique individual appointment descriptors and came to a consensus rating of each. This process resulted in what we refer to as Treatment appointments and Non-therapy appointments, where Treatment appointments encompass all appointments that were consensually agreed to potentially be part of routine counseling (and represent a "dose" of this treatment), while Non-therapy appointments included other appointments at the counseling center. Importantly, standard intake evaluations and initial assessments were included as Treatment appointments because they were deemed part of the routine course of therapy, whereas case management, crisis triage meetings, medication check-ups, and phone calls were considered Non-therapy appointments.

The total number of therapy sessions attended is an important consideration in studies of effectiveness. A stringent test of treatment effectiveness includes all clients who attended treatment at any time, regardless of their total therapy dose, while a more liberal but arguably equitable test might exclude clients who failed to attend a sufficient number of sessions. This would be more equitable for comparisons to other trials, especially RCTs of individual psychotherapy that commonly include only clients who complete a minimum number of sessions (or the complete treatment course) in analyses. We elected to include clients if they attended at least one Treatment session for two reasons: First, because there is some evidence that in applied psychotherapy settings, treatment length is dependent on progress and not strictly vice-versa (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009; Barkham et al., 2006), and second, because this provides a more stringent test

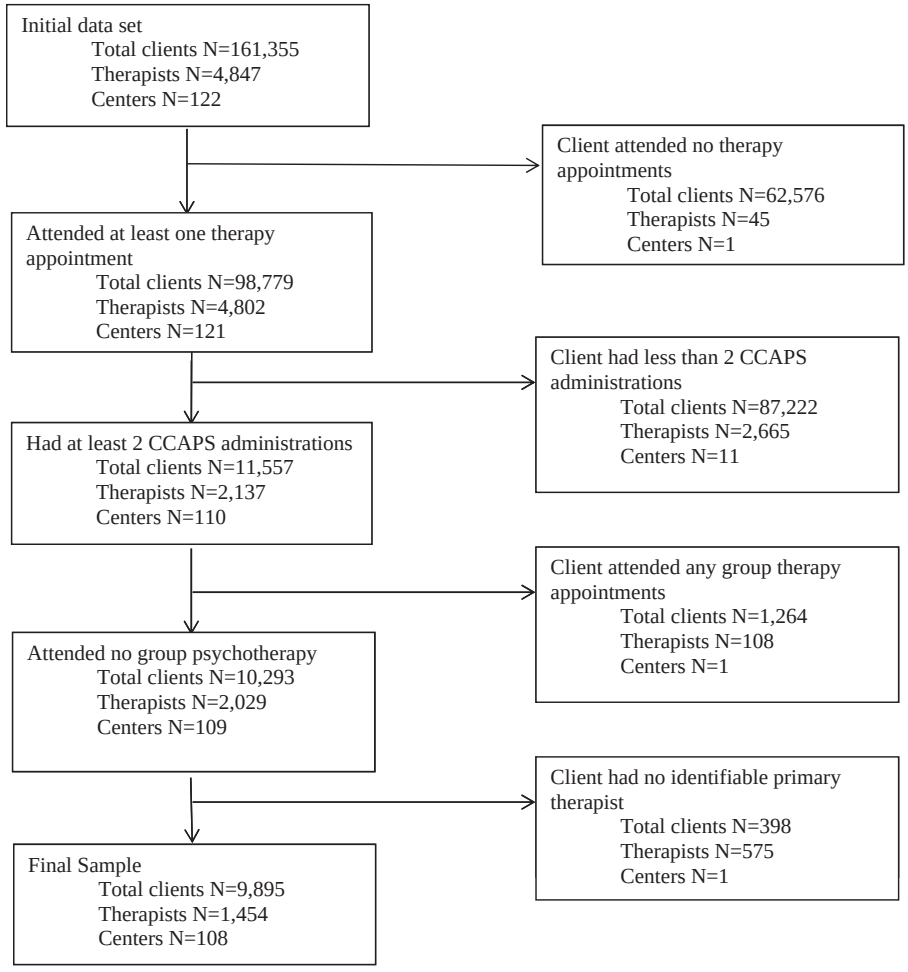

Figure 1. Data reduction.

of routine treatment services in UCCs (as clients who discontinued treatment before the completion of a course of therapy were included). We further excluded any clients who attended more Non-therapy sessions than Treatment sessions, and all clients who attended any group appointments, in order to include only those clients engaged in routine individual face-to-face counseling as a primary or exclusive treatment within the counseling center. Additionally, we excluded clients who did not have an identifiable "primary" therapist: a therapist who saw them for at least 50% of their Treatment sessions. This decision was made in order to identify clients who had an opportunity to develop a therapeutic relationship with a single therapy provider, and exclude those who presented for multiple sessions with multiple providers (e.g., due to crisis), but might not be in treatment with any given individual.

Several additional inclusion and exclusion rules were required due to the diversity of services across centers. We required that the client completed at least two assessments of symptomatology (see measures below): one within 14 days of the first attended appointment and one within 14 days of the last attended appointment. We refer to these as pre- and post-treatment observations. These 14-day windows allow for the inclusion of centers that administer the first assessment of symptoms prior to any services, as well as the centers that administer the same assessment less frequently than every session. Our expectation is that if these expanded pre- and post-treatment windows add any systematic bias to the analyses, it should reduce the apparent effect of treatment since this includes cases whose final observed outcome measure could be two weeks before the end of treatment (and therefore missing some treatment dose) or up to two weeks after treatment ends (at which time clients may have lost some of their treatment gains). Following Minami et al. (2009) we defined "courses" of therapy within this multi-year data as those that included no more than 90 days between sessions. For clients with multiple courses of therapy, we only included the first course of treatment.

The final data set included 9,895 clients seen by 1,454 therapists at 108 UCCs. The average age of clients was 22.7 years (SD = 5.3). The majority, 6,257 (67.5%) were female, 2,951 (31.9%) were male, 21 identified as transgender and 28 preferred not to answer. The vast majority—6,320, or 71.9% —identified as Caucasian, with 687 (7.8%) identifying as Hispanic/Latino/a, 649 (7.4%) as Black/African American, 472 (5.3%) as Asian American/Asian, 318 (3.6%) as multi-racial, with no other category accounting for more than 1% of the total sample.

Treatment dose as indicated by the number of Treatment appointments was highly variable, ranging from 1 to 86 with a mean of 7.10 sessions (SD = 5.39, median = 6, mode = 5). However, the vast majority (97.0%) of clients had 20 or fewer sessions. Of these, the mean number of attended Treatment appointments was 6.49 (SD = 3.89), with a median of 6 and a mode of 5. It should be noted that the structure of the data is highly complex: occasions nested within clients, crossed with or nested within therapists, nested within centers. The variability associated with these higher levels (therapist and center) was not modeled in the present study. The variance associated with centers was quite small in the data: the largest intraclass correlation across the CCAPS-34 subscales showed less than 3% of the variability related to centers, with other subscales below 2%. This is generally considered very small or negligible (e.g., Tabachnick & Fiddell, 2007). Therapist effects were not accounted for in this sample for primarily analytical reasons: requiring a minimum number of clients per therapist would drastically reduce the overall number of clients in the sample. Also, as discussed in the data reduction steps, many clients saw multiple therapists, making the analysis of variance components more computationally intensive and less easily interpreted. Additionally, since the study did not compare treatment groups within the TAU sample, one of the primary benefits of accounting for higher-level effects—namely reducing Type I error between groups—was not pertinent.

Measures

Counseling center assessment of psychological symptoms (CCAPS; Locke et al., 2011, 2012). The CCAPS is a multidimensional instrument designed to assess several problems common to students seeking services at counseling centers. It has two versions: 62 items (CCAPS-62) and 34 items (CCAPS-34), with the latter developed to facilitate repeated assessment of treatment progress and outcome. The CCAPS instruments are administered according to each UCCs' policies and procedures. Frequently, the CCAPS-62 is administered at intake, and the CCAPS-34 is used for repeated administrations. Later administrations are often at every session, but other schedules (e.g., every two sessions, once per month) are also used at different UCCs. For the purpose of this study, all administrations of the CCAPS were scored as the 34-item version, which has seven subscales: Depression, Generalized Anxiety, Social Anxiety, Hostility, Academic Distress, Eating Concerns, and Alcohol Use. A general distress measure, the Distress Index (DI), is also used in practice which is an aggregate score based on 20 items across all but the Eating Concerns and Alcohol Use subscales, and as such provides a measure of general distress and negative affect across symptom groups (Nordberg et al., 2016).

All subscales have been shown to have good internal consistency (Cronbach's alpha ranging from .83-.89), criterion validity (strong correlations with established measures of similar constructs), and discriminant validity (low correlations with unrelated constructs; Locke et al., 2011). Further, the subscales of the CCAPS (except for Alcohol Use) have been shown to predict treatment utilization, and all subscales with clear diagnostic analogs have been found to be substantially elevated in diagnosed clients compared to clients not given a diagnosis (McAleavey et al., 2012).

In practice two cut points are used for each subscale, theoretically dividing each into three regions (consistent with other measures common in practice, e.g., Ronk et al., 2013). As developed by McAleavey et al. (2012), the lower cut point is based on Jacobson and Truax's (1991) criterion "c" cut point, which is the distributional midpoint between a nonclinical and a clinical group. The higher cut point, for five of the subscales, is the result of receiver operating characteristic curve analyses of the subscales predicting diagnoses among clients in counseling, with the negative predictive power of the high cut point being extremely high (above .95), while the positive predictive power is relatively low (less than .5; see McAleavey et al., 2012, for more information). The low cut point can be interpreted to separate individuals with little relevant psychopathology from those who might have some mild distress, and the higher cut point separates those with mild distress from those who are most likely to have a specific clinical concern in the relevant symptom domain (CCMH, 2012). High scores on the CCAPS-34 are not diagnostic, but diagnoses are unlikely at lower levels of the subscale scores.

Standardized data set (SDS; CCMH, 2012). The SDS is a flexibly administered standardized questionnaire for demographics and treatment history of counseling center clients. Individual UCCs can determine which questions to present to their clients, in what order, and can add additional questions before, during, and after the SDS administration, though CCMH recommends a set pattern. Because of this, missing data on the SDS is frequently due to a UCC's administration choice rather than the client's unwillingness to answer. Generally, UCCs administer the SDS to incoming clients prior to their first session and rarely more than once.

Missing Data

Given the nature of this data, there are unique challenges to missing data. Since the data are collected directly from an EMR system, if there was a session or a CCAPS, the EMR would capture it and thus, no data loss is believed to have occurred at the point of entry. Since each center can set their own policy on measure administration schedules, some missing data occur when individuals are still in treatment but are not assessed. In these cases, the missing data will be related to center-level decisions, not treatment or client characteristics directly. However, we determined that modeling or imputing these missing observations would be both analytically challenging (given the range of missing data patterns, schedules of administration, and limited numbers of observations (<3) for most clients) and would be less representative of the observations made by administrators evaluating individual UCCs where weekly CCAPS administration is not possible. Therefore, we elected to treat the observed values as they were taken, so the analyses depicted here are accurate to observed values, but may misestimate the effectiveness of routine treatments due to this variability. Additionally, there are two other points where missing data may occur: First, incomplete CCAPS responses (i.e., missing item responses), and second, dropout. Incomplete CCAPS administrations are processed in the same manner as Locke et al. (2011) which amounts to identifying a small number of administrations where fewer than half of the items on a subscale were completed and treating those administrations as invalid. Remaining missing items are negligible in the data (< 1% missing for all items). Dropout is not modeled or treated specially in this data as this is part of the comparison in question. The use of data here amounts to last-observation-carried-forward in a treatment study (when such clients could be included in the study). For most analyses in this paper, listwise or pairwise deletion are considered appropriate given the nature of the data being as close to population-level statistics as possible.

Data Analysis

All analyses were conducted using R (R Core Team, 2013).

RCI calculations. The Jacobson and Truax (1991) RCI for each subscale of the CCAPS-34 are published in the CCAPS Technical Manual (CCMH, 2012). As an initial comparison, we calculated the percent of clients in each subsample who achieved reliable improvement on the appropriate RCI-based criterion. Though it is clear that some of the clients likely deteriorated, we did not calculate this separately because this value does not have a clear parallel in the other methods used in the study.

Clinical trials benchmarking. Procedures used for clinical trials benchmarking closely replicated those described in detail by Minami et al. (2007, 2009; Minami, Serlin, et al., 2008). Those studies described the development and use of one set of clinical benchmarks: adult outpatient psychotherapy for MDD. These authors developed three primary sets of benchmarks depending on the type of outcome measure used, and recommend applying different benchmarks based on the available data. Specifically, they separated measures according to specificity (how symptom-specific versus general a measure is) and reactivity (self-report symptom measures are expected to change less than provider reported measures). Given that the subscales of the CCAPS are symptom-specific and self-report, the appropriate comparison measures are low-reactivity, high-specificity (LR-HS) measures. In the case of the Depression CCAPS subscale, we adopted Minami et al.'s (2007) LR-HS benchmark for MDD treatments, which is based on analyses of the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). The CCAPS' general distress measure, the DI, in contrast, is a low-reactivity, low-specificity (LR-LS) measure, and for this measure we adopted their LR-LS MDD benchmark, consistent with Minami, Wampold, et al.'s (2008) analysis of outpatient psychotherapy.

However, for the other CCAPS subscales, which are all LR-HS measures, there were no benchmarks available, and therefore, required the calculation of new benchmarks. Minami et al. (2007) conducted a meta-analytic review of the research literature on depression treatments to derive their MDD

benchmarks. Their review led to the inclusion of 35 total studies, though several specific benchmarks in their study were based on many fewer studies (e.g., their LR-LS intent-to-treat (ITT) benchmark had $k = 4$ studies). We adopted a different strategy from Minami et al.: rather than conducting a meta-analysis de novo for each subscale, we determined to rely on recently published meta-analyses of treatments for specific disorders. The strength of this method is that the benchmarks adhere closely to a large body of relevant literature in each domain, as selected by separate groups of expert researchers.

Accordingly, we identified meta-analyses of psychotherapy treatments for each of the subscales of the CCAPS other than depression. We consulted with experts in their respective fields to ensure that the meta-analyses found were representative. In some cases, when a single meta-analysis was not comprehensive (such as when the authors stated that the literature review began recently or excluded clients relevant to our study), we included multiple meta-analyses per subscale. Our final selection of meta-analyses included psychotherapeutic treatments for generalized anxiety (Gould, Otto, Pollack, & Yap, 1997; Mitte, 2005), social anxiety (Acarturk, Cuijpers, Van Straten, & De Graaf, 2009), anger management (DiGuiseppe & Tafrate, 2003; Del Vecchio & O'Leary, 2004), and eating disorders (Thompson-Brenner, Glass, & Westen, 2003; Vocks et al., 2010). No relevant meta-analyses could be identified for academic distress, as this is not a typical focus of psychotherapy. Once we had identified the meta-analyses, we located as many source studies therein by online resources and contacting authors. For inclusion in benchmarks, we required that the treatments were face-to-face individual psychosocial interventions, and selected outcome measures that matched the CCAPS subscales in reactivity and specificity (that is, we identified self-report, symptom-specific outcome measures within each study of each meta-analysis). When multiple qualifying measures were reported, we used only one measure per study, with priority given first to the most common instrument across studies (to minimize method variance), and then to the measure that the authors reported as their primary outcome. From these we attempted to follow Minami, Serlin, et al.'s (2008) recommendations to identify separate completer and ITT benchmarks for only bona fide treatments. However, the number of studies fully reporting ITT results was deemed too small to be useful. Therefore, we used values for completer samples, which were reported universally. These studies were used as our TE benchmark sources. Waitlist conditions were used as NH benchmarks.

For the Alcohol Use subscale, we identified two potentially relevant meta-analyses (Dutra et al., 2008; Imel, Wampold, Miller, & Fleming, 2008). However, measurement exclusion criteria yielded a relatively limited sample of includable studies from these two meta-analyses. Specifically, the majority of the primary outcomes in alcohol and substance use treatment studies are biological or count-based measures rather than self-report scales of symptoms, and these types of outcome do not represent comparable benchmarks for the CCAPS Alcohol Use subscale. Therefore, in order to augment the studies included for the benchmark, we conducted a new review on PsycINFO for the Alcohol Use Disorders Identification Test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993), which is a widely used self-report measure of alcohol use highly correlated with the CCAPS-34 Alcohol Use subscale (Locke et al., 2012). In the search, 2853 studies using the AUDIT were identified. The abstracts were then reviewed for English language, treatment provision, randomization, and AUDIT administration at pre- and post-treatment, and these inclusion criteria were used to search for RCTs of individual psychotherapies for alcohol use disorders that also used a LR-HS measure of alcohol use at pre- and post-treatment. However, no new studies were identified through this process that met all of our criteria for inclusion. We continued with the analyses from the first search, but we consider the results related to alcohol use to be preliminary, and report them here for completeness. Descriptive information regarding each meta-analysis is in Table I, with additional information regarding individual source studies available in the Supplementary Materials.

Once we had identified source trials for each subscale, we implemented Minami et al.'s (2007, 2009) methods for developing TE and NH benchmarks. This is essentially a two-step procedure, requiring first that an unbiased estimate of effect size be calculated for each condition of each study, and then these effect size estimates are aggregated across studies into benchmarks. Unbiased effect sizes (d_i) were computed following Minami, Serlin, et al.'s (2008) Equation (1) for individual sample i:

$$d_i = \left(1 - \frac{3}{4n_i - 5}\right) \frac{M_{i,\text{post}} - M_{i,\text{pre}}}{\text{SD}_{i,\text{pre}}},$$

where n_i is the sample size, $M_{i,\text{post}}$ is the post-treatment mean of the measure, $M_{i,\text{pre}}$ is the pre-treatment mean of the measure, and $\text{SD}_{i,\text{pre}}$ is the pre-treatment standard deviation of the measure. When more than one LR-HS measure was reported for a given study,

Table I. Description of benchmark studies.

CCAPS-34 Subscale	Disorders or treatment target for benchmarks	Meta-analyses	Number of studies included	Number of discrete bona fide TE samples (Number of discrete NH samples)	Total *N* in TE benchmark (Total *N* in NH benchmarks)	Most commonly used instruments
Generalized Anxiety	Generalized anxiety disorder	Gould et al. (1997); Mitte (2005)	15	27 (7)	395 (83)	STAI Trait Anxiety
Social Anxiety	Social phobia	Acarturk et al. (2009)	15	30 (12)	530 (177)	Fear of Negative Evaluation
Eating Concerns	Anorexia nervosa, bulimia, binge eating disorder	Thompson-Brenner et al. (2003); Vocks et al. (2010)	44	82 (9)	2179 (180)	Eating Attitudes Questionnaire-26; Eating Disorders Inventory
Hostility	Anger management treatments	Del Vecchio and O'Leary (2004); DiGiuseppe and Tafrate (2003)	33	54 (10)	1191 (80)	STAXI Trait Anger Scale
Alcohol Use	Alcohol use disorders, Substance use disorders	Dutra et al. (2008); Imel et al. (2008)	6	17 (1)	2322 (69)	The Drinker Inventory of Consequences

Notes: Summary of the meta-analyses and papers used in deriving the benchmarks for this study. TE: Treatment Efficacy; NH: Natural History; STAI: State-Trait Anxiety Inventory; STAXI: State-Trait Anger Expression Inventory. For information regarding the Depression benchmark, see Minami et al. (2007). Additional information can be found in the Supplementary Materials.

only one was included in our analyses, as determined by a count of the most common measures for each symptom type. The most common outcome measures for each symptom type are included in Table I.

The variance of d_i is given by Minami, Serlin, et al.'s (2008) Equation (2):

$$\hat{\sigma}^2_{d(i)} = \frac{2(1-r_i)}{n_i} + \frac{d_i^2}{2n_i},$$

where r_i is the correlation between pre- and post-treatment outcome measures. In practice, this value must be estimated; Minami, Serlin, et al. (2008) suggest a moderate value of $r = .5$ be used for outcome measures. The benchmark study effect sizes are then aggregated into a single benchmark effect size d_B using the formula:

$$d_B = \frac{\sum_i d_i/\hat{\sigma}^2_{d(i)}}{\sum_i 1/\hat{\sigma}^2_{d(i)}}.$$

This process is repeated, separately, for the waitlist conditions of the identified benchmark studies, yielding TE benchmark and NH benchmarks separately. The effect size for the TAU sample and the variance of that estimate are then calculated using similar formulae (Minami et al., 2009, Equations (4) and (5), respectively).

The benchmarks are then compared to the TAU effectiveness estimate, using a range-null hypothesis test. Specifically, a margin of "clinical equivalence" of $d \pm 0.2$ is adopted to allow inferences of whether the TAU and benchmark samples differ meaningfully rather than whether they are differentiable from each other at all. This results (in the case of TE benchmarks) in Minami, Serlin, et al.'s (2008) Equations (6) and (7):

$$H_0: \delta_D \leq \delta_{B(\mathrm{TE})} - \Delta,$$
$$H_1: \delta_D > \delta_{B(\mathrm{TE})} - \Delta,$$

where δ_D is the true effect size of the TAU sample, $\delta_{B(\mathrm{TE})}$ is the true effect size of the TE benchmark, and Δ is the clinical equivalence value (defined a priori as 0.2).[2] The test follows a noncentral t distribution with a noncentrality parameter $\lambda_{\mathrm{TE}} = \sqrt{N}(\delta_{B(\mathrm{TE})} - \Delta)$ and $N-1$ degrees of freedom, with N being the TAU sample size. Interpretation is aided by the calculation of a critical value $d_{\mathrm{CV(TE)}}$ in Minami, Serlin, et al.'s (2008) Equation (8). If the TAU effect size estimate exceeds the critical value for the TE benchmark, the null hypothesis is rejected and TAU effectiveness is inferred to be not meaningfully less than the clinical trials benchmark. The inverse is true

for the NH benchmark, though the method is the same: in that case, the test of interest examines whether the TAU sample is meaningfully more effective than the NH benchmark, rather than less. We conducted benchmark comparisons for each of the available subscales of the CCAPS-34 on three different samples of the TAU data: (1) the overall sample, (2) only people who scored above the low cut point at pre-treatment, and (3) only those who scored above the high cut point at pre-treatment.

Normative comparisons. The method of normative comparisons (both as proposed by Kendall et al., 1999; van Wieringen & Cribbie, 2014) requires that the population be determined to be distressed at the start of treatment, so the data for these analyses exclude individuals whose scores were below the low cut point for each subscale at pre-treatment.

The primary concern in establishing norms is to select a group of people who can be expected to adequately represent the range of normal functioning (Kendall et al., 1999). The CCAPS instruments have been used in a large, national survey of college students conducted by NASPA (for more detailed description, see, McAleavey et al., 2012). Briefly, this survey included over 19,000 students who completed the CCAPS-62 and a measure of treatment history online. As noted by Kendall et al. (1999), it is important to not exclude all individuals in the normative sample who are currently experiencing symptoms of interest, so as not to create a "supernormal," non-representative sample. For this reason, we adopted the group of participants who reported no current treatment for psychological concern as our normative group (referred to as the "No Treatment" group by McAleavey et al., 2012).

Normative comparisons were conducted in R using functions provided by R. A. Cribbie (retrieved online from https://github.com/cribbie/equivalencetests 24 February 2017). Specifically, the samples were trimmed for outliers and extreme values, and then Windsorized means and variances were calculated for each sample. The tests of equivalency were two complementary *t*-tests of the difference between group means as defined in van Wieringen and Cribbie (2014), in which the null hypotheses are that the difference between means is not less than the range of normative equivalence. The normative range of equivalence was set at 1 SD from the mean of the normative sample. As discussed above, the S-Y test is robust to violations of non-normality, and examinations of normal Q-Q plots of our data suggested that meaningful non-normality was present. Two comparisons were run for each subscale: the first for all TAU participants whose first CCAPS-34 subscale scores exceeded the low cut point, and the second for the subset of these whose first subscale score also exceeded the high cut point.

Results

Descriptive information regarding change in the clinical sample can be found in Table II. Across all subscales, greater change magnitude was observed among the subsets with greater severity at pre-treatment. This is consistent with previous findings, and may partially be due to regression to the mean and spontaneous recovery (Barkham et al., 2008, 2012). The percent reliably improved ranged from 8% for the Alcohol Use subscale in the total sample to 49% for Depression in the Elevated subsample, again with more substantial results occurring in the more severely distressed subsamples, across all subscales. In general, the highly distressed subsample on each subscale showed reliable improvement rates consistent with previous research, though Depression showed the greatest rates, and almost double the improvement rate of some other subscales. In addition, we calculated the percent of the sample reporting at least one reliable improvement across all of the subscales: 52.6% of the Overall sample had at least one subscale with an RCI change. That is, more than half of all clients in the study showed reliable improvement on at least one subscale, from pre- to post-treatment.

The change benchmarking results are reported in Table III, including the benchmark target values calculated based on the meta-analyses studies for each of the CCAPS-34 subscales. In terms of the comparison to NH benchmarks, notably, of 18 effect sizes, only two (eating concerns and alcohol use, both for the overall sample) were not clinically superior to the NH benchmark. That is, almost every test demonstrated that clients reported greater change during treatment than would be expected without treatment, regardless of their pre-treatment severity state. However, the results for clinical equivalence to the TE benchmark were less positive. Although five of six comparisons among the highest distress group did find equivalence, we did not find any results in the overall sample that were clinically equivalent to the TE benchmark. That is, receiving services in clinical routine of counseling centers was not associated with equivalent improvement to TE, across all domains of symptoms. However, for individuals with an identified highly elevated area of distress, counseling was clinically equivalent in all areas

Table II. Pre–post descriptive analyses in the routine treatment data set.

CCAPS-34 Subscale	Subsample	*N*	Pre-treatment Mean	Pre-treatment SD	Post-treatment mean	Post-treatment SD	Percent reliably improved (%)
Depression	Overall	9895	1.73	1.00	1.09	0.90	28.78
	Low cut point	7014	2.23	0.72	1.34	0.90	40.58
	Elevated cut point	4896	2.58	0.56	1.52	0.93	49.47
Generalized Anxiety	Overall	9895	1.89	0.98	1.44	0.92	20.37
	Low cut point	7081	2.36	0.72	1.70	0.88	28.22
	Elevated cut point	4152	2.84	0.52	1.99	0.88	36.97
Social Anxiety	Overall	9895	1.94	1.01	1.65	0.95	11.90
	Low cut point	3098	2.65	0.62	2.14	0.81	18.67
	Elevated cut point	5754	3.13	0.42	2.49	0.78	24.05
Eating Concerns	Overall	9895	0.98	1.15	0.83	1.05	9.46
	Low cut point	3323	2.39	0.85	1.77	1.11	28.14
	Elevated cut point	2701	2.63	0.75	1.94	1.10	29.91
Hostility	Overall	9895	0.98	0.88	0.67	0.71	11.99
	Low cut point	6627	1.67	0.71	1.04	0.76	23.50
	Elevated cut point	2974	2.04	0.62	1.23	0.79	34.62
Alcohol Use	Overall	9894	0.70	0.94	0.53	0.79	8.03
	Low cut point	3776	1.68	0.84	1.10	0.92	21.03
	Elevated cut point	2493	2.09	0.74	1.35	0.95	31.85

Notes: Results of the observed TAU at UCCs on each subscale of the CCAPS-34. Each row represents a subsample of the overall sample, for whom the subscale of interest is reported.

except Alcohol Use. Results for the moderately distressed groups were intermediate, with some subscales, such as Generalized Anxiety, Eating Concerns, and Hostility showing clinical equivalence to the TE benchmark, but not for others, such as Depression, Social Anxiety, and Alcohol Use.

The results from the normative comparisons presented in Table IV show the opposite pattern: None of the seven subscales showed a return to the normal range for clients who entered treatment above the high cut score. However, the other subsets (which made smaller overall changes as observed in the RCI and benchmark analyses) did tend to be equivalent to the normative sample by post-treatment across all subscales of the CCAPS-34. That is, the method of normative comparison suggested treatment effectiveness in the opposite pattern of CSC and treatment benchmarking in this sample, likely as a function of a smaller distance needed to change to meet this criterion in less-distressed clients.

Discussion

In this study, we attempted to put the effectiveness of counseling into a meaningful and valid frame, by assessing several different symptom domains and by using two methods to measure effectiveness: change-based and end-state analyses. With the goal of obtaining findings that could be generalized to similar environments, populations, and services, we conducted this study in a large sample of college counseling clients across 108 clinical settings and imposed very few exclusion criteria beyond what was necessary to identify clients who entered individual psychotherapy. This study is among the broadest assessments of in vivo mental health treatments of which we are aware, with implications for the assessment of effectiveness and for the measurement of effectiveness in future studies.

Perhaps the most important of our findings is that clients who enter counseling with a defined area of distress show a large amount of change, clinically equivalent to the amount observed in gold-standard RCTs for several psychological conditions. These findings should provide strong reason for confidence in the individual counseling services offered at UCCs, and are consistent with results from Minami et al. (2009) and Minami, Wampold, et al. (2008) who reported similar outcomes from comparisons of outpatient services. It is worth noting that a recent study by Reese, Duncan, Bohanske, Owen, and Minami (2014), which did not include normative end-state comparisons but reported similar results for benchmarking of depression in routine care,

Table III. Clinical efficacy benchmarking results.

		Benchmarks				TAU			
CCAPS-34 Subscale	Subsample	TE benchmark	NH benchmark	TE critical value	NH critical value	*N*	Unbiased pre–post effect size	Clinically equivalent to TE benchmark?	Clinically superior to NH benchmark?
Depression	Overall	−1.706	−0.371	−1.530	−0.589	9895	−0.639	No	Yes
	Low cut point			−1.535	−0.592	7014	−1.240	No	Yes
	Elevated cut point			−1.541	−0.597	4896	−1.896	Yes	Yes
Generalized Anxiety	Overall	−0.890	0.026	−0.782	−0.154	9895	−0.465	No	Yes
	Low cut point			−0.784	−0.156	7081	−0.906	Yes	Yes
	Elevated cut point			−0.792	−0.163	4152	−1.632	Yes	Yes
Social Anxiety	Overall	−1.051	−0.039	−0.900	−0.256	9895	−0.296	No	Yes
	Low cut point			−0.906	−0.261	3098	−0.823	No	Yes
	Elevated cut point			−0.915	−0.269	5754	−1.515	Yes	Yes
Eating Concerns	Overall	−0.823	−0.196	−0.721	−0.453	9895	−0.132	No	No
	Low cut point			−0.735	−0.466	3323	−0.735	Yes	Yes
	Elevated cut point			−0.738	−0.469	2701	−0.926	Yes	Yes
Hostility	Overall	−0.910	−0.023	−0.735	−0.192	9895	−0.356	No	Yes
	Low cut point			−0.743	−0.199	6627	−0.889	Yes	Yes
	Elevated cut point			−0.749	−0.204	2974	−1.307	Yes	Yes
Alcohol Use	Overall	−1.288	−0.038	−1.054	−0.255	9894	−0.183	No	No
	Low cut point			−1.067	−0.265	3776	−0.684	No	Yes
	Elevated cut point			−1.075	−0.272	2493	−0.995	No	Yes

Notes: Comparisons between our observed sample from UCCs (TAU) and the benchmarks derived from the meta-analyses. TE and NH benchmarks are the target values calculated based on the studies used to derive the benchmarks. Bolded "Yes" entries in the last two columns represent either clinical equivalence to the benchmarking studies, or being clinically superior to the NH benchmark, each of which would indicate that routine treatments are providing clinically useful services. Alcohol Use results are considered tentative, due to a small number of samples, especially in the NH benchmark ($k = 1$). NH: Natural history; TAU: Treatment as usual; TE: Treatment efficacy.

concluded that outpatient psychotherapy treatment was "likely effective" (p. 738). The results of the present study may similarly suggest that counseling in routine practice is likely effective at reducing symptoms of several psychological disorders (MDD, generalized anxiety disorder, social phobia, eating disorders), but does not result in a return to normative functioning, on average, for those clients who start treatment with the most severe difficulties. The fact that there were only minor differences on benchmarking results as a function of symptom cluster should be noted as well. Though the absolute amount of change (noted in Table II) varied considerably across subscales, the conclusions across TE and NH benchmarks were more alike than not. With the possible and tentative exception of Alcohol Use, routine counseling was associated with similar conclusions across subscales of the CCAPS-34. Adding another source of confidence toward the effectiveness of individual services delivered in counseling centers, we did not identify particular symptoms that fail to respond to routine treatment.

It is also worth identifying the benchmarks used in this study as an additional contribution. As shown in Table III, there were sizeable differences across diagnostic group in both TE and NH benchmarks. The TE benchmarks ranged from $d = -0.823$ (Eating Disorders) to $d = -1.706$ (Depression), illustrating that in RCTs these populations experience considerably different pre–post changes. Perhaps more surprisingly, the NH benchmarks also showed meaningful variability across diagnosis, ranging from $d = -0.371$ (Depression) to $d = 0.026$ (Generalized Anxiety

Table IV. Normative comparisons results.

		Normative sample			Clinical sample post-treatment			
CCAPS Subscale	Subsample	Mean	SD	Trimmed mean	Mean	SD	Trimmed Mean	S-Y equivalent?
Depression	Overall	0.722	0.783	0.530	1.088	0.897	0.956	**Yes**
	Low cut point				1.342	0.896	1.259	**Yes**
	Elevated cut point				1.516	0.927	1.464	No
Generalized Anxiety	Overall	1.025	0.798	0.908	1.436	0.921	1.357	**Yes**
	Low cut point				1.705	0.884	1.666	**Yes**
	Elevated cut point				1.991	0.883	1.998	No
Social Anxiety	Overall	1.422	0.893	1.343	1.647	0.946	1.602	**Yes**
	Low cut point				2.141	0.814	2.134	**Yes**
	Elevated cut point				2.488	0.779	2.522	No
Eating Concerns	Overall	0.938	1.035	0.674	0.830	1.054	0.507	**Yes**
	Low cut point				1.768	1.113	1.722	No
	Elevated cut point				1.936	1.098	1.916	No
Hostility	Overall	0.589	0.671	0.418	0.669	0.715	0.504	**Yes**
	Low cut point				1.035	0.760	0.938	**Yes**
	Elevated cut point				1.232	0.794	1.159	No
Alcohol Use	Overall	0.658	0.882	0.373	0.528	0.788	0.255	**Yes**
	Low cut point				1.105	0.920	0.977	**Yes**
	Elevated cut point				1.353	0.954	1.275	No
Academic Distress	Overall	1.198	0.921	1.076	1.522	1.044	1.438	**Yes**
	Low cut point				1.872	0.999	1.843	**Yes**
	Elevated cut point				2.287	0.995	2.338	No

Notes: Comparisons between our observed sample from UCCs at post-treatment with the normative values derived from a large national dataset. Bolded "Yes" entries in the last column represent end-state distributions equivalent to the normative sample according to the Schuirmann-Yuen test, which would indicate that routine treatments are providing clinically useful services. All decisions are based on the combination of two *t*-tests, and all were $p < .001$.

Disorder), indicating that some symptoms are considerably more likely to resolve spontaneously than others. Certain symptoms (generalized anxiety disorder, social phobia, anger, and alcohol use) appear to have almost no change during a waitlist period, while depression and eating disorders seem to undergo small changes over time. This illustrates the importance of using different benchmarks for different symptom groups: if only one benchmark is used (say for depression), some symptoms would have appeared to be clinically inferior to TE benchmarks. The benchmarks presented here can be used, amended and improved by future research, and possibly serve as basic targets for ongoing quality assurance programs in outpatient mental health.

Another important finding of the study is that the subsamples of clients who had initial CCAPS-34 subscale scores above the Elevated cut point showed the greatest change across all subscales. This suggests that therapists are able to target specific symptom clusters when they are present. These individuals are considered to be the most clinically similar to individuals with identified clinical disorders and problems on a subscale, and are therefore likely to be the most similar to participants in RCTs for these disorders; perhaps this is the best comparison group for benchmarking. However, there are at least two measurement difficulties that make this finding difficult to interpret by itself. First, it is necessarily the case that higher cut scores are associated with lower pre-treatment SDs due to restricted range at pre-treatment observation, which subsequently makes effect size measures appear larger (because the pre-treatment SD is the preferred denominator in calculating effect sizes). In addition, regression to the mean is expected to play a greater role with the higher-severity clients, since many individuals' high initial scores may be due to chance fluctuations and measurement error, so that larger changes may not reflect truly larger effects of therapy. Importantly, both of these measurement difficulties are true of any RCT employing a cut score inclusion criteria (or, by extension, diagnosis) as well as the previous literature on benchmarking effect sizes in routine practice. As such, these concerns warrant future examination.

These complications are one of the main reasons that including a normative comparison approach to routine treatment outcome assessment is advisable. End-state comparisons, unlike comparisons of change, will not be preferentially affected by regression to the mean and artificially shrunken pre-treatment SDs. Instead, an opposing bias may be

present: it is easier for a group to be in the normal range at post-treatment if they started treatment closer to that threshold, and increasingly greater change is required to observe a return to normative function for more distressed individuals. This increasing cost works in opposition to regression to the mean. The use of end-state analyses led to a third important finding, which is that highly distressed clients did not, as a group, return to "normal" symptomatic levels by the end of treatment. Combined with the results of change-based benchmarking, this finding demonstrates the utility of using in tandem two complementary methods of measuring effectiveness: in this case the conclusions are actually opposed to one another and the end-state analyses suggest that treatment-as-usual can be improved.

Regardless of methods for defining effectiveness, we believe that using a meaningful cut score at baseline is necessary when broadly examining routine treatment outcomes. A meaningful cut score is one that helps identify clients who would be most likely to have an identifiable disorder or problem. On the CCAPS-34 and other screening measures, high cut points most effectively provide meaningful discrimination. The multidimensional nature of the CCAPS-34 means that many individuals who seek treatment will not report distress on certain subscales. Therefore, even though the overall data set used here is informative of the total population entering treatment, it is not clearly informative regarding people who enter treatment for a specific problem. Only individuals who begin treatment above a certain meaningful point should be considered to even potentially—let alone likely—require treatment for that symptom cluster. In a controlled trial, participants are selected based on diagnostic criteria, often in conjunction with a cut score on some standardized instrument, and therefore only a subset of the clients in our high cut point groups would be eligible for an RCT, and almost no one below this cut score would be included in such a study. Accordingly, analyses involving the high cut point can be considered conservative in comparison to RCTs; the CCAPS-34 high cut point likely captures many individuals who have elevated distress in a particular domain without necessarily also having a diagnosis, and therefore may not have even been in treatment for the outcome being assessed here.

It is very important to keep in mind the number of sessions attended in this study. Evidence-based psychotherapy treatments are often tested as 6–20 session treatments (e.g., Cuijpers et al., 2014). The benchmark papers used here mostly fell within this 6–20 session range, and with averages between 8 and 16 sessions per person for each benchmark, representing a substantially greater dose of treatment than our routine treatment sample, which had an average of 6.49 of attended sessions. We chose a very small number (1) of minimum treatment sessions, and if a dose-effect model of psychotherapy exists, this would bias the estimated treatment effect downward. Given the substantial variability between patients, therapists, and UCCs, a more thorough examination of treatment length in this data would require further evaluation of other variables (e.g., session limit per center, session frequency, dropout, and external referrals). Though such a study would be valuable, it would go beyond the scope of this paper, which was intended to assess treatment effectiveness across symptom type and severity in the total sample of clients seeking psychotherapy. Though we cannot conclude this with certainty, one possibility is that the average treatment response would be greater if clients attended a larger number of sessions, which in many UCCs is not possible due to session limits and clinical volume.

With this in mind and based on the results discussed above, we propose that this study suggests that even though routine treatment seems to be effective in terms of facilitating improvement in symptoms for the most distressed individuals (clinically equivalent change to RCTs), important quality improvements to routine treatments can still be made. Specifically, further focus on individuals with highly elevated scores at intake may help bring the post-treatment functioning to within normal range. Providing symptom-targeted evidence-based interventions to these clients in particular may make these improvements. Other clinical and administrative changes, such as extending the number of sessions of therapy offered, working with multiple treatment modalities (e.g., pharmacological, social work), and/or providing appropriate referrals may further improve care.

Limitations

There are some important limitations to this study. The first, as mentioned above, is regression to the mean. This is clearly relevant when examining different cut points, as higher distressed individuals may be expected to make larger gains based only on regression to the mean, rather than psychotherapy effects. This is especially a problem for benchmarking methods which are exclusively focused on the magnitude of change, though incorporating a NH control group should account for true regression to the mean and spontaneous recovery.

Another limitation worth mentioning is the adequacy of clinical trials benchmarks: We present our

review of meta-analyses as an example of development of benchmarks, and other authors and future researchers could produce different benchmarks than we established for the present study. The methods used here, namely to rely on previous meta-analyses for a targeted problem, have several advantages, however. First, it can be relatively easily adapted to other problems, and in many cases may be possible to extract directly from meta-analyses that report within-group effects. Second, it represents the closest approximation of the state of research in each field, so long as the meta-analyses are adequately recent. And further, though we did not control for study quality, doing so frequently brings the efficacy estimates lower than the estimates used here. For example, Cuijpers, Smit, Bohlmeijer, Hollon, and Andersson (2010) suggested that controlling for study quality would transform the mean (between-group) effect size of psychotherapy for Depression from $d = .67$–$.42$. In conjunction with our use of completer rather than ITT samples, this suggests that the TE benchmarks reported here might be on the high end of what could be expected, and represent a higher standard for routine care evaluation. Nevertheless, there are countless differences between samples collected in routine settings and those from clinical trials for specified disorders, and benchmarking is at best a highly specified analog method, not a true experiment.

Another limitation is that we purposefully did not examine moderators of effectiveness beyond symptom type and initial severity. Other potential moderators would include center characteristics, such as session limits and types of treatments offered, co-occurring service utilization, treatment duration (as discussed above), and numerous client characteristics. Additionally, despite the implications of therapist effects in assessing treatment effectiveness (Baldwin & Imel, 2013), therapist effects were also not included in the present study for several reasons. First, few therapists saw enough clients to produce a reliable estimate of therapist effectiveness. Second, many clients saw multiple therapists during their treatment, which is highly typical of clinical practice in UCCs, but complicates analysis of therapists. Finally, one major aim of this study is to compare effect sizes to past research which has not accounted for differences between therapists, generally. These moderators should be examined in future research.

Finally, in this study we cannot be certain exactly what treatments were being administered. In fact, we do not know that all of the clients in the study truly were in treatment at all, only that they attended a session that can reasonably be expected to be part of individual psychotherapy and were not engaged in substantial additional treatments within their UCC. This is an inherent limitation of data collection on a large scale in an applied setting: though we included many clients, we know relatively little about each one. In all cases, we attempted to err on the side of caution, making decisions that would not inflate a treatment effect. It is possible that these choices served to actually reduce the apparent effectiveness of routine psychotherapy in this study, and less restrictive studies may have different results.

Conclusions

Examining and reporting the overall effectiveness of routine care is an important and sometimes overlooked part of improving care available to clients. In this study, we found that across several domains of symptoms, clients who enter treatment distressed undergo equivalent improvement in symptom severity to clients in RCTs. However, those same clients do not on average return to within normal limits following treatment. In combination, this suggests that individual psychological treatments administered in UCCs are effective but could still be improved, especially for highly distressed clients. The mechanisms that may most benefit these clients are currently unknown, but may include longer or more intense psychotherapy, additional types of treatment, or other solutions. Moreover, this demonstrates that the method used to assess effectiveness matters a great deal, as does the choice of cut scores: researchers can reach completely different conclusions, depending on choices they make. Future research would benefit from more explicitly identifying definitions of effectiveness a priori and incorporating both change-based and end-state assessments.

Supplemental data

Supplemental data for this article can be accessed here. https://doi.org//10.1080/10503307.2017.1395921

Notes

[1] Several methods exist to account for initial severity and other case-specific variables in such analyses, from the use of cut scores to determine eligibility from routine samples (as in the case of most previous research), propensity scores (Lutz, Schiefele, Wucherpfennig, Rubel, & Stulz, 2016), and machine-learning algorithms (Kraus et al., 2016). A thorough discussion of these issues is beyond the current scope of this paper.

[2] An alternative method used by Reese et al. (2014) and others proposes an interval of 10% of the effect size rather than a fixed value of 0.2. This method will be more conservative with regard to TE benchmarks and more liberal with regard to NH benchmarks than the fixed value of 0.2, as long as the observed

effect size is below $d = 2$. We elected to use the fixed value method because it has an absolute interpretation rather than a relative one. In addition, calculations using the relative value method resulted in only one interpretive change of the 36 comparisons.

ORCID

Andrew A. McAleavey http://orcid.org/0000-0001-5986-2033
Henry Xiao http://orcid.org/0000-0001-5620-1096
Jeffrey A. Hayes http://orcid.org/0000-0003-4716-1015

References

Acarturk, C., Cuijpers, P., Van Straten, A., & De Graaf, R. (2009). Psychological treatment of social anxiety disorder: A meta-analysis. *Psychological Medicine*, *39*, 241–254.

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.

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). Hoboken, NJ: Wiley.

Barkham, M., Connell, J., Stiles, W. B., Miles, J. N. V., Margison, F., Evans, C., & Mellor-Clark, J. (2006). Dose–effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology*, *74*, 160–167.

Barkham, M., Stiles, W. B., Connell, J., & Mellor-Clark, J. (2012). Psychological treatment outcomes in routine NHS services: What do we mean by treatment effectiveness? *Psychology and Psychotherapy: Theory, Research and Practice*, *85*, 1–16.

Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M., … Angus, L. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology*, *47*, 397–415.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Center for Collegiate Mental Health. (2012). *CCAPS 2012 technical manual*. University Park, PA.

Cribbie, R. A., & Arpin-Cribbie, C. A. (2009). Evaluating clinical significance through equivalence testing: Extending the normative comparisons approach. *Psychotherapy Research*, *19*, 677–686. doi:10.1080/10503300902926554

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., & Reynolds, C. F. (2013). The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: A meta-analysis of direct comparisons. *World Psychiatry*, *12*(2), 137–148.

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *The British Journal of Psychiatry*, *196*, 173–178.

Cuijpers, P., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., & Coyne, J. (2014). Comparison of psychotherapies for adult depression to pill placebo control groups: A meta-analysis. *Psychological Medicine*, *44*, 685–695.

Del Vecchio, T., & O'Leary, K. D. (2004). Effectiveness of anger treatments for specific anger problems: A meta-analytic review. *Clinical Psychology Review*, *24*, 15–34.

DiGiuseppe, R., & Tafrate, R. C. (2003). Anger treatment for adults: A meta-analytic review. *Clinical Psychology: Science and Practice*, *10*, 70–84. doi:10.1093/clipsy.10.1.70

Dutra, L., Stathopoulou, G., Basden, S. L., Leyro, T. M., Powers, M. B., & Otto, M. W. (2008). A meta-analytic review of psychosocial interventions for substance use disorders. *American Journal of Psychiatry*, *165*, 179–187.

Forand, N. R., Evans, S., Haglin, D., & Fishman, B. (2011). Cognitive-behavioral therapy in practice: Treatment delivered by trainees at an outpatient clinic is clinically effective. *Behavior Therapy*, *42*, 612–623.

Gould, R. A., Otto, M. W., Pollack, M. H., & Yap, L. (1997). Cognitive behavioral and pharmacological treatment of generalized anxiety disorder: A preliminary meta-analysis. *Behavior Therapy*, *28*, 285–305.

Hunsley, J., & Lee, C. M. (2007). Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. *Professional Psychology: Research and Practice*, *38*, 21–33.

Imel, Z. E., Wampold, B. E., Miller, S. D., & Fleming, R. R. (2008). Distinctions without a difference: Direct comparisons of psychotherapies for alcohol use disorders. *Psychology of Addictive Behaviors*, *22*, 533–543.

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.

Kendall, P. C., & Grove, W. M. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment*, *10*, 147–158.

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

Kraus, D. R., Bentley, J. H., Alexander, P. C., Boswell, J. F., Constantino, M. J., Baxter, E. E., & Castonguay, L. G. (2016). Predicting therapist effectiveness from their own practice-based evidence. *Journal of Consulting and Clinical Psychology*, *84*(6), 473–483.

Lambert, M. J., Hatfield, D. R., Vermeersch, D. A., Burlingame, G. M., Reisinger, C. W., & Brown, G. S. (2003). *Administration and scoring manual for the OQ-30.1*. East Setauket, NY: American Professional Credentialing Services.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53–70.

Locke, B. D., Buzolitz, J. S., Lei, P., Boswell, J. F., McAleavey, A. A., Sevig, T. D., … Hayes, J. A. (2011). Development of the counseling center assessment of psychological symptoms-62 (CCAPS-62). *Journal of Counseling Psychology*, *58*(1), 97–109.

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P., Hayes, J. A., Castonguay, L. G., … Lin, Y. (2012). Development and initial validation of the counseling center assessment of psychological symptoms-34. *Measurement and Evaluation in Counseling and Development*, *45*(3), 151–169.

Lutz, W., Schiefele, A.-K., Wucherpfennig, F., Rubel, J., & Stulz, N. (2016). Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice. *Journal of Affective Disorders*, *189*, 150–158.

Mangardich, H., & Cribbie, R. A. (2014). Assessing clinical significance using robust normative comparisons. *Psychotherapy Research*, *25*, 239–248. Advance online publication.

McAleavey, A. A., Lockard, A. J., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2015). Building a practice research network: Obstacles faced and lessons learned at the center for collegiate mental health. *Psychotherapy Research*, *25*(1), 134–151. doi:10.1080/10503307.2014.883652

McAleavey, A. A., Nordberg, S. S., Kraus, D., & Castonguay, L. G. (2012). Errors in treatment outcome monitoring: Implications for real-world psychotherapy. *Canadian Psychology*, *53*, 105–114.

McEvoy, P. M., & Nathan, P. (2007). Effectiveness of cognitive behavior therapy for diagnostically heterogeneous groups: A benchmarking study. *Journal of Consulting and Clinical Psychology*, *75*, 344–350.

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*(2), 404–409.

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

Mitte, K. (2005). Meta-analysis of cognitive-behavioral treatments for generalized anxiety disorder: A comparison with pharmacotherapy. *Psychological Bulletin*, *131*, 785–795.

Nasiakos, G., Cribbie, R. A., & Arpin-Cribbie, C. A. (2010). Equivalence based tests of clinical significance: Assessing treatments for depression. *Psychotherapy Research*, *20*, 647–656. doi:10.1080/10503307.2010.501039

Nordberg, S. S., McAleavey, A. A., Duszak, E., Locke, B. D., Hayes, J. A., & Castonguay, L. G. (2016). The counseling center assessment of psychological symptoms distress index: A pragmatic exploration of general factors to enhance a multidimensional scale. *Counselling Psychology Quarterly*, 1–16. doi:10.1080/09515070.2016.1202809

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Reese, R. J., Duncan, B. L., Bohanske, R. T., Owen, J. J., & Minami, T. (2014). Benchmarking outcomes in a public behavioral health setting: Feedback as a quality improvement strategy. *Journal of Consulting and Clinical Psychology*, *82*, 731–742.

Ronk, F. R., Korman, J. R., Hooke, G. R., & Page, A. C. (2013). Assessing clinical significance of treatment outcomes using the DASS-21. *Psychological Assessment*, *25*, 1103–1110. doi:10.1037/a0033100

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the alcohol Use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption–II. *Addiction*, *88*, 791–804. doi:10.1111/j.1360-0443.1993.tb02093.x

Sheldrick, R. C., Kendall, P. C., & Heimberg, R. G. (2001). The clinical significance of treatments: A comparison of three treatments for conduct disordered children. *Clinical Psychology: Science & Practice*, *8*, 418–430.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Pearson.

Thompson-Brenner, H., Glass, S., & Westen, D. (2003). A multidimensional meta-analysis of psychotherapy for bulimia nervosa. *Clinical Psychology: Science and Practice*, *10*, 269–287.

Vocks, S., Tuschen-Caffier, B., Pietrowsky, R., Rustenbach, S. J., Kersting, A., & Herpertz, S. (2010). Meta-analysis of the effectiveness of psychological and pharmacological treatments for binge eating disorder. *International Journal of Eating Disorders*, *43*, 205–217. doi:10.1002/eat.20696

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*(2), 231–239.

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*(2), 299–310.

van Wieringen, K., & Cribbie, R. A. (2014). Evaluating clinical significance: Incorporating robust statistics with normative comparison tests. *British Journal of Mathematical and Statistical Psychology*, *67*, 213–230.