Treatment Integrity of Studies That Compare Short-Term Psychodynamic Psychotherapy With Cognitive-Behavior Therapy

Sunil S. Bhar and Aaron T. Beck, Department of Psychiatry, University of Pennsylvania

In their meta-analysis, Leichsenring, Rabung, and Leibing (2004) found that short-term psychoanalytic psychotherapy (STPP) and cognitive-behavior therapy (CBT) were equally efficacious. However, they did not examine whether the treatments were delivered as intended. In our review of nine studies comparing STPP and CBT, we found that most did not adequately implement treatment integrity procedures. Therefore, CBT and STPP are equally efficacious is therefore based on studies without adequate procedures to monitor treatment adherence or therapist competence. The neglect of treatment integrity procedures in these studies leaves open the possibility that the absence of comparative treatment effects may be due to the manner in which the treatments were operationalized.

Key words: cognitive-behavior therapy, meta-analysis, short-term psychodynamic psychotherapy, treatment integrity. [Clin Psychol Sci Prac 16: 370–378, 2009]

The notion that all psychotherapies produce similar outcomes has been extensively debated. Some researchers assert that there is little evidence for superior treatment effects for any form of psychotherapy for any disorder (e.g., Wampold et al., 1997), a position dubbed the “dodo bird verdict” (Luborsky, Singer, & Luborsky, 1975). Such assertions are supported by early meta-analyses showing that different psychotherapy models produce similar outcomes (Shapiro & Shapiro, 1982; Smith & Glass, 1977). However, more recently, researchers have argued that many comparative studies cannot be interpreted unambiguously because researchers did not adequately implement procedures for evaluating treatment integrity (Kazdin, 1986; Nezu & Nezu, 2008).

Treatment integrity refers to the extent to which treatments are implemented as intended (Perepletchikova & Kazdin, 2005) and is commonly conceptualized as encompassing three related aspects: (a) therapist treatment adherence, that is, the degree to which the therapist utilizes prescribed procedures, (b) therapist competence, that is, the level of the therapist’s skill and judgment, and (c) treatment differentiation, that is, whether treatments differ from each other along critical dimensions (Perepletchikova, Treat, & Kazdin, 2007; Waltz, Addis, Koerner, & Jacobson, 1993). Treatment integrity has important implications for drawing inferences about the comparative efficacy of different treatments. Without adequate implementation of treatment integrity in comparative studies of psychotherapy, it is difficult to establish whether the conclusion about the superiority of one treatment over another (or conversely,
the lack of observed differences between treatments) is the result of poorly operationalized treatments. A treatment could be found to be superior to another because one treatment was delivered well, while the other was poorly implemented. Conversely, the absence of treatment effects may be due to poor adherence, competence, or differentiation, resulting in unintended overlap between the treatments (Nezu & Nezu, 2008; Perepletchikova et al., 2007; Waltz et al., 1993).

Psychotherapy research reports have been criticized for inadequately attending to treatment integrity procedures (Nezu & Nezu, 2008). In a recent review of 202 psychosocial treatments, Perepletchikova et al. (2007) found adequate reporting of adherence and competence procedures for only 8.9% and 1.5% of treatments, respectively. The authors acknowledged that these patterns of results may reflect incomplete reporting, rather than the omission of procedures utilized for implementing or detecting integrity. However, they and other authors (e.g., Borrelli et al., 2005) suggest that the lack of attention placed on treatment integrity in such reports reduces readers’ confidence in interpreting outcomes of psychotherapy research.

To what extent have randomized controlled trials (RCTs) of short-term psychodynamic psychotherapy (STPP) and cognitive-behavior therapy (CBT) included protocols to establish and assess treatment integrity? Over the past 40 years, CBT and STPP have emerged as empirically supported forms of treatment for a range of psychiatric disorders (Abbass, Hancock, Henderson, & Kisely, 2006; Butler, Chapman, Forman, & Beck, 2006) and the subject of numerous comparative research studies and meta-analyses (Anderson & Lambert, 1995; Crits-Christoph, 1992; Leichsenring, 2001; Leichsenring, Rabung, & Leibing, 2004; Svarthberg & Stiles, 1991). Equivocal findings have emerged about the relative efficacy of the two treatment approaches. Some meta-analyses have concluded that CBT is more efficacious than STPP (Grawe, Donati, & Bernauer, 1994; Svarthberg & Stiles, 1991), while others have found equivalent outcomes for both treatments (Anderson & Lambert, 1995; Crits-Christoph, 1992; Leichsenring, 2001; Leichsenring et al., 2004). As far as we know, the quality of treatment integrity protocols across such RCTs has not been formally examined; therefore, it remains uncertain whether the outcomes of these analyses reflect substantive differences or similarities between the treatments or the inadequate implementation of treatment protocols.

In part to redress the lack of attention paid to the quality of treatment integrity procedures in RCTs, Leichsenring et al. (2004) conducted a meta-analysis of the efficacy of STPP. Compared with previous meta-analyses, Leichsenring et al. screened out RCTs that compared treatments that could not be clearly discriminated from each other. For instance, they excluded studies of interpersonal psychotherapy (IPT) because it was not clear how IPT differed from either STPP (Markowitz, Svarthberg, & Swartz, 1998) or CBT (Ablon & Jones, 2002). They also excluded RCTs that did not use treatment manuals or a manual-like guide.

The aim of their analysis was to compare the efficacy of STPP with that of other psychotherapies along three outcome dimensions: target problems (e.g., anxiety levels in treatment of patients with anxiety disorders), general psychiatric symptoms (e.g., depression levels in treatment of patients with personality disorder), and social functioning. They concluded that STPP was equally efficacious compared with other psychotherapies, including CBT, when aggregating results across a range of psychiatric disorders such as depressive disorders, eating disorders, substance use disorders, and personality disorders. Of the nine published studies reviewed in their analysis, seven found no difference between treatment outcomes. Reminiscent of the dodo bird verdict (Luborsky et al., 1975), Leichsenring et al. (2004) equated their results as “consistent with those of the meta-analysis of Wampold et al. (1997), who did not find differences between bona-fide methods of psychotherapy” (p. 1213).

However, Leichsenring et al. (2004) did not systematically examine the quality of treatment integrity procedures in their cohort of studies. Therefore, it is possible that the null findings in their meta-analysis were due to poor treatment adherence, differentiation, or variable levels of competence in therapists, resulting in unintended overlap between treatment conditions (Kazdin, 1986; Waltz et al., 1993). The extent to which these studies addressed treatment integrity remains to be empirically examined.

The definition of what comprises adequate procedures for monitoring treatment integrity has been...
operationalized by Perepletchikova et al. (2007). According to their approach, the quality of treatment integrity procedures can be assessed (using a rating scale called the Implementation of Treatment Integrity Procedures Scale, ITIPS) against multiple recommendations proposed in the literature for establishing, assessing, evaluating, and reporting treatment adherence and therapist competence (Perepletchikova et al., 2007). First, according to Perepletchikova et al. (2007), researchers should establish procedures for monitoring adherence and competence. Adherence should be described as the degree of utilization of procedures, which are to be specified in a treatment manual. If this definition is extended to include the avoidance of prescribed procedures, then treatment differentiation can also be monitored by adherence scales. An assessment of treatment adherence is sufficient for determining whether treatments are different from each other if treatment manuals or adherence measures include a list of procedures to be avoided by the therapist (Perepletchikova et al., 2007). Competence should be defined as the level of skills and judgment shown by the therapist in delivering treatment, rather than on the basis of the therapist’s previous experience.

Second, the assessment of adherence, differentiation, and competence should ideally be assessed using direct observations (e.g., review of audio recordings), rather than indirect evaluations (e.g., self-reports by therapists or patients). Although more costly than indirect methods, direct observations by raters are less subject to distortion due to self-presentation biases and poor recollections (Perepletchikova et al., 2007). Ratings of treatment adherence, differentiation, and competence should employ measures that are valid and reliable. Ideally, psychometric properties of these measures should also be presented in the article or referenced.

Third, the procedures for assessing treatment integrity should be evaluated for biases. Given that therapists may adhere more closely to the treatment protocols if they are aware of being observed, treatment integrity procedures should deliberately control for such reactivity (e.g., all sessions are recorded or observed, therapists are interviewed at random times, etc.). Finally, numerical data that are informative of the levels of adherence, differentiation, and competency in the study should be reported. Such reporting procedures are expected to describe the extent to which each treatment was delivered as intended.

The aim of the current study was to investigate the extent to which these recommendations were adequately addressed in the nine published studies comparing CBT and STPP that were included in Leichsenring and colleagues’ (2004) meta-analysis. We limited our review to those studies as they represent the most methodological, rigorous comparisons of CBT and STPP to date. Using the framework proposed by Perepletchikova et al. (2007), we reviewed the quality of the treatment integrity procedures for each of the studies.

**METHOD**

**Selection of Studies**

The nine published studies comparing CBT and STPP reviewed by Leichsenring et al. (2004) were selected for the current review (see Table 1).

**Measure**

The ITIPS, a structured rating scale for treatment integrity (Perepletchikova et al., 2007), was used to assess treatment adherence and therapist competence in each study. The ITIPS contains 22 items that measure four domains of treatment integrity—that is, the manner in which treatment integrity is established (six items), assessed (seven items), evaluated (five items), and reported (four items). Items are also grouped into subscales that address procedures relating specifically to the two main aspects of treatment integrity: treatment adherence (six items; adherence subscale) and therapist competence (six items; competence subscale). A total of six subscales is derived from the ITIPS items (four domains and two aspects). In order to examine if the evaluation of treatment adherence was sufficient for determining treatment differentiation, we reviewed each article for information on whether their treatment manuals or adherence measures included a list of proscribed behaviors. Each item on the ITIPS was rated on a four-point scale ranging from 1 to 4, where higher ratings indicated more adequate implementation of integrity procedures. The scoring of the scales followed rules and procedures outlined by Perepletchikova et al. (2007). On each item, a score of 1 or 2 was assumed to reflect inadequate implementation of integrity procedures; a score of 3
indicated that implementation approached adequacy; and a score of 4 designated adequate implementation of integrity procedures. Following Perepletchikova et al. (2007), studies scoring in the bottom half of the score range were classified as inadequately implementing integrity procedures. Studies scoring in the top 25% range of the total score were classified as adequately implementing integrity procedures. Studies scoring between these two ranges were classified as approaching adequacy. For example, studies that scored less than 13 on the competency subscale were classified as inadequately addressing therapist competency. Studies that received a score of 18 or higher on that subscale were classified as adequately addressing competency. Studies scoring between 13 and 18 were classified as approaching adequate implementation of procedures for addressing therapist competency. Classification ranges for each of the subscales are provided in Perepletchikova et al. (2007, p. 833).

Perepletchikova et al. (2007) reported high levels of internal consistency and inter-rater agreement. For the four domain and two integrity subscales, Cronbach’s alphas ranged from .76 to .93 (M = .84, SD = 0.7); inter-rater agreement (T index) for the ITIPS was 0.89.

Procedures

The nine studies were rated following procedures outlined in Perepletchikova et al. (2007) and guidelines in their rating manual (available on http://www.treatmentintegrity.com). Two undergraduate students (both female, 21 and 22 years of age, both White) who were blinded to the aims of the study were trained as raters. Two 1-hour sessions spread over 2 weeks were implemented in the training of raters. No articles used for training purposes were included in the sample. Following the training, the raters scored each article separately. When authors of articles referred to outside sources for further information pertaining to treatment integrity procedures, these sources were consulted to make informed ratings (see Table 1 for supplemental articles). References were consulted for five of the nine articles. Where there were inconsistencies between the ratings, the mean ratings on each item were employed. Of the 128 ratings made per rater on the ITIPS, 77 were inconsistent, yielding an inter-rater agreement (T) index of .48. Positive values of T indicate that rater agreement is greater than chance (Tinsley & Weiss, 2000). Cronbach’s alphas for the subscales ranged from .85 to .94.

Unlike Perepletchikova et al. (2007), who rated each treatment for treatment integrity, we rated studies because we found identical treatment integrity procedures were adopted for all treatment conditions in a study. For example, when direct methods of observation were used to assess treatment integrity levels in one treatment condition, these methods were also

### Table 1. Adequacy of implementation of treatment integrity procedures

<table>
<thead>
<tr>
<th>Study</th>
<th>Total treatment integrity</th>
<th>Establishing treatment integrity</th>
<th>Assessing treatment integrity</th>
<th>Evaluating treatment integrity</th>
<th>Reporting treatment integrity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score Classified</td>
<td>Score Classified</td>
<td>Score Classified</td>
<td>Score Classified</td>
<td>Score Classified</td>
</tr>
<tr>
<td>Woody et al. (1990)*</td>
<td>33.5 IA</td>
<td>13.5 AA</td>
<td>9.0 IA</td>
<td>5.5 IA</td>
<td>5.5 IA</td>
</tr>
<tr>
<td>Fairburn et al. (1986)</td>
<td>33.0 IA</td>
<td>9.5 IA</td>
<td>12.0 IA</td>
<td>5.0 IA</td>
<td>6.5 IA</td>
</tr>
<tr>
<td>Thompson et al. (1987)</td>
<td>45.5 AA</td>
<td>16.0 AA</td>
<td>14.0 IA</td>
<td>6.5 IA</td>
<td>9.0 AA</td>
</tr>
<tr>
<td>Brom et al. (1989)</td>
<td>31.5 IA</td>
<td>12.0 IA</td>
<td>8.5 IA</td>
<td>5.0 IA</td>
<td>6.0 IA</td>
</tr>
<tr>
<td>Garner et al. (1993)</td>
<td>42.0 IA</td>
<td>16.5 AA</td>
<td>13.0 IA</td>
<td>5.0 IA</td>
<td>7.5 IA</td>
</tr>
<tr>
<td>Shapiro et al. (1994)*</td>
<td>62.0 AA</td>
<td>23.0 AD</td>
<td>21.0 AD</td>
<td>7.0 IA</td>
<td>11.0 AA</td>
</tr>
<tr>
<td>Crits-Christoph et al. (1999)*</td>
<td>78.5 AD</td>
<td>24.0 AD</td>
<td>26.0 AD</td>
<td>15.0 AA</td>
<td>13.5 AD</td>
</tr>
<tr>
<td>Cooper et al. (2003)*</td>
<td>41.8 IA</td>
<td>14.5 AA</td>
<td>12.3 IA</td>
<td>6.5 IA</td>
<td>8.5 AA</td>
</tr>
<tr>
<td>Svartberg et al. (2004)*</td>
<td>61.0 AA</td>
<td>21.5 AD</td>
<td>21.5 AD</td>
<td>8.0 IA</td>
<td>10.0 AA</td>
</tr>
<tr>
<td>Average across studies (SD)</td>
<td>47.6 (16.1) AA</td>
<td>16.7 (5.1) AA</td>
<td>15.3 (6.1) AA</td>
<td>7.1 (3.2) IA</td>
<td>8.6 (2.6) AA</td>
</tr>
</tbody>
</table>

Note: Rating classifications are provided by Perepletchikova et al. (2007) according to score ranges on the Implementation of Treatment Integrity Procedures Scale (ITIPS). IA, inadequate implementation; AA, approaching adequacy; AD, adequate.

*List of supplemental articles consulted (Barber et al., 1996; Barber et al., 2004; Crits-Christoph et al., 1998, 2001; Gaston et al., 1998; Shapiro & Startup, 1992; Silove et al., 1990; Startup & Shapiro, 1993; Woody et al., 1983, 1986).

*Reliability and validity data are provided for the full 66-item version of the Therapist Rating Scale but not the abbreviated 30-item version (Silove et al., 1990).
used to assess integrity levels in the other treatment conditions of a study (e.g., Fairburn, Kirk, O’Connor, & Cooper, 1986; Garner et al., 1993; Thompson, Gallagher, & Breckenridge, 1987).

RESULTS

As seen in Table 1, the ratings for the implementation of treatment integrity procedures (total score on the ITIPS) were 56% inadequate, 33% approaching adequacy, and 11% adequate. Treatment integrity was established 22% inadequately, 44% with approaching adequacy, and 33% adequately. Treatment integrity was assessed 67% inadequately, 0% with approaching adequacy, and 33% adequately. Treatment integrity was evaluated 89% inadequately, 11% with approaching adequacy, and 0% adequately. Treatment integrity levels were reported 44% inadequately, 44% with approaching adequacy, and 11% adequately.

As shown in Table 2, procedures specific to treatment adherence were implemented across studies 67% inadequately, 33% with approaching adequacy, and 22% adequately (mean adherence rating = 13.1, SD = 4.8; mean classification: “approaching adequacy”). All studies were found to refer to and define treatment adherence with 78% employing direct methods instead of, or in addition to, indirect methods for assessing adherence. Treatment adherence was assessed with valid and reliable measures, or while controlling for therapist reactivity in 44% of studies. Numerical information about adherence levels was reported in 56% of studies.

No study indicated that their treatment manuals included a list of procedures to be avoided. Only four studies (44%) employed adherence scales that included items measuring proscribed behaviors (Cooper, Murray, Wilson, & Romanik, 2003; Crits-Christoph et al., 1999; Shapiro et al., 1994; Svarberg, Stiles, & Seltzer, 2004). The remaining five studies either did not assess treatment differentiation (Brom, Kleber, & Defares, 1989; Woody, Luborsky, McElhaney, & O’Brien, 1990) or provide criteria for differentiating between treatments (Fairburn et al., 1986; Garner et al., 1993; Thompson et al., 1987).

Table 2 also shows that therapist competence procedures were implemented 66.7% inadequately, 22.2% with approaching adequacy, and 11.1% adequately (mean competence rating = 10.1, SD = 5.0; mean classification: “inadequate”). Therapist competence was defined and measured directly in 44.4% of studies. Therapist competence was assessed while controlling for therapist reactivity in 22% of studies. Numerical information about competence level was measured by valid and reliable measures and reported in 11% of studies.

DISCUSSION

The extent to which different psychotherapies are equally efficacious has been extensively debated. Leichsenring et al. (2004) propose that STPP and CBT are equally efficacious across a range of psychiatric disorders. They acknowledge that their findings are based on a limited number of studies that aggregate effects across a broad range of psychiatric disorders. In this article, we add a further caveat to their conclusion: Their findings are based on studies, most of which do not adequately implement treatment integrity procedures.

According to our review of the studies, no study met criteria for adequately implementing all four domains of treatment integrity measured by the ITIPS. Of the nine studies, only three were judged to adequately establish and assess treatment integrity, one to adequately report numerical data that was informative of the levels of adherence, and none to sufficiently ensure the accuracy of the obtained integrity data. In summary, most studies did not satisfy criteria for adequately implementing treatment integrity procedures.

Procedures specific to treatment adherence were implemented less than adequately in seven of the nine studies. The following studies exemplify problems implementing procedures for addressing treatment adherence. Brom et al. (1989) implemented supervisory sessions to ensure adherence to procedures of three types of treatments of posttraumatic stress disorder (PTSD) but did not measure nor report the extent to which the therapists were adherent to the treatment protocols. They admit that “the similarity of results in the three treatment conditions may be due to similarities in the behavior of therapists, which we did not measure directly…” (p. 610). Likewise, Woody et al. (1983, 1990) found no differences between the efficacy of CBT and supportive expressive therapy for drug use in opiate addicts. However, they did not provide...
information about whether therapists adhered to or deviated from the respective treatment models, thus also leaving open the possibility that similarity in treatment outcomes was because of an unintended overlap between the treatments. Further, Cooper et al. (2003) monitored levels of adherence to treatment protocols; however, they assessed adherence indirectly, through patients’ report of therapist behavior, rather than on the basis of observations of the sessions.

Table 2. Adequacy of implementation of treatment integrity procedures specifically for treatment adherence and for therapist competence

<table>
<thead>
<tr>
<th>Study</th>
<th>Score</th>
<th>Classified</th>
<th>Definition established</th>
<th>Assessed using direct or indirect methods</th>
<th>Assessed with valid measure</th>
<th>Assessed with reliable measure</th>
<th>Evaluated and controlled for reactivity</th>
<th>Reported informative data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment adherence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woody et al. (1990)a</td>
<td>8.0</td>
<td>IA</td>
<td>Yes</td>
<td>Direct</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Fairburn et al. (1986)</td>
<td>11.0</td>
<td>IA</td>
<td>Yes</td>
<td>Direct</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Thompson et al. (1987)</td>
<td>12.0</td>
<td>IA</td>
<td>Yes</td>
<td>Direct</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Brom et al. (1989)</td>
<td>8.5</td>
<td>IA</td>
<td>Yes</td>
<td>Indirect</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Garner et al. (1993)</td>
<td>13.0</td>
<td>AA</td>
<td>Yes</td>
<td>Both</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Shapiro et al. (1994)a</td>
<td>19.0</td>
<td>AD</td>
<td>Yes</td>
<td>Direct</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Crits-Christoph et al. (1999)a</td>
<td>20.0</td>
<td>AD</td>
<td>Yes</td>
<td>Direct</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cooper et al. (2003)</td>
<td>14.0</td>
<td>AA</td>
<td>Yes</td>
<td>Indirect</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Svarberg et al. (2004)a</td>
<td>17.0</td>
<td>AA</td>
<td>Yes</td>
<td>Direct</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Therapist competence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woody et al. (1990)a</td>
<td>6.0</td>
<td>IA</td>
<td>No</td>
<td>Indirect</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Fairburn et al. (1986)</td>
<td>7.0</td>
<td>IA</td>
<td>No</td>
<td>Neither</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Thompson et al. (1987)</td>
<td>12.5</td>
<td>IA</td>
<td>Yes</td>
<td>Direct</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Brom et al. (1989)</td>
<td>6.0</td>
<td>IA</td>
<td>No</td>
<td>Neither</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Garner et al. (1993)</td>
<td>6.0</td>
<td>IA</td>
<td>No</td>
<td>Neither</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Shapiro et al. (1994)a</td>
<td>13.5</td>
<td>AA</td>
<td>Yes</td>
<td>Direct</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Crits-Christoph et al. (1999)a</td>
<td>20.0</td>
<td>AD</td>
<td>Yes</td>
<td>Direct</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cooper et al. (2003)</td>
<td>6.0</td>
<td>IA</td>
<td>No</td>
<td>Neither</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Svarberg et al. (2004)a</td>
<td>13.5</td>
<td>AA</td>
<td>Yes</td>
<td>Direct</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Note: Rating classifications are provided by Perepletchikova et al. (2007) according to score ranges on the Implementation of Treatment Integrity Procedures Scale (ITIPS). IA, inadequate implementation (ratings between 6 and 12); AA, approaching adequacy (13–18); AD, adequate (19–24).

*Treatment integrity was assessed using a list of proscribed tasks as defined by treatment adherence measures (Cooper et al., 2003; Crits-Christoph et al., 1999; Shapiro et al., 1994; Svarberg et al., 2004). Thus, in these four studies, the measurement of treatment adherence extended to treatment differentiation. In two studies, treatment differentiation was not assessed at all (Brom et al., 1989; Woody et al., 1990). In three studies, no criteria were provided to explain how treatments were discriminated from each other by raters (Fairburn et al., 1986; Garner et al., 1993; Thompson et al., 1987). Therefore, in these five studies, we do not know the extent to which the therapists incorporated activities from other interventions not specified in a protocol.

Therapist competence procedures were found to be implemented less than adequately in eight of the nine studies. The following studies highlight some of the reasons for this finding. In Woody and colleagues’ study, therapists of each modality were selected as “best for this population” (Woody et al., 1983, p. 641) by the supervisor for each therapy and were supervised weekly or biweekly throughout the study. However, no competency checks were reported. Therefore, it was not clear whether either treatment was delivered competently by the therapists. Further, Cooper et al. (2003) employed specialists and nonspecialists across three treatment approaches (CBT, STPP, and nondirective counseling) and found no difference in outcomes between the three approaches. The authors did
not assess the extent to which the therapists competently delivered the treatments, thus leaving ambiguous the question of whether experts and nonexperts differed in terms of their competency levels and the impact of such competency on outcomes. Similarly, two other studies (Fairburn et al., 1986; Garner et al., 1993) did not assess the competency of therapists for each treatment condition (CBT, STPP) and therefore cannot rule out the extent to which the superiority of CBT to STPP is attributed to differences in competency levels between CBT and STPP therapists.

Only one of the nine studies reviewed was found to comprehensively describe procedures for establishing, assessing, evaluating, and reporting treatment integrity (Crits-Christoph et al., 1999). The authors of this study used manual-guided protocols and provided extensive training to therapists to establish treatment integrity (see Crits-Christoph et al., 1998). They selected only those clinicians who demonstrated ability to deliver treatments as intended to participate in the trial (Barber, Foltz, Crits-Christoph, & Chittams, 2004). They employed experts to assess audio recordings of a random set of therapy sessions; these raters were, in turn, evaluated for inter-rater reliability (see Barber et al., 2004). Ratings were made using reliable and valid measures of adherence and competence. Treatment sessions from each therapy condition were audited for whether therapists utilized prescribed behaviors and avoided proscribed behaviors. Data on treatment integrity were reported for each treatment condition (Barber et al., 2004).

Our results are reflective of past research on the implementation of treatment integrity procedures (Nezu & Nezu, 2008; Perepletchikova et al., 2007). Similar to observations by Perepletchikova et al. (2007), we found that only a small proportion of studies adequately attended to either treatment adherence or therapist competence. Moreover, mirroring their findings, we found that fewer studies adequately monitored therapist competence than adherence aspects of treatment integrity. For example, while all nine studies in our sample monitored treatment adherence, only four monitored therapist competence. Perhaps, as suggested by some authors, therapist competence is more complex to operationalize and expensive to rate compared to treatment adherence (Perepletchikova et al., 2007; Waltz et al., 1993).

Two limitations of this study are important to note when considering these findings. First, the ratings of each study do not reflect whether the therapist adhered to the treatment protocol or were competent in delivering the treatments. Rather, the ratings refer to the extent to which the researchers adequately established protocols to check deviations from treatment integrity. Therefore, even though a study is rated low on the ITIPS, it is possible that the study treatments were delivered competently and according to the treatment protocols. Conversely, studies that are rated high on the ITIPS may not have delivered the treatments as intended. Thus, the documentation of treatment integrity does not indicate levels of adherence or competence for each treatment condition. However, as noted in Perepletchikova et al. (2007), regardless of the level of treatment integrity, neglecting to report procedures for implementing or tracking treatment integrity makes conclusions about treatment efficacy ambiguous, as readers cannot know the extent to which results of the studies are affected by low treatment integrity.

Second, the rating instrument used in this study was relatively new. Although the ITIPS is associated with high levels of internal consistency and inter-rater agreement, it has not yet been validated with respect to other rating schemes of treatment integrity (e.g., Borrelli et al., 2005). The cutoff scores used in this study to classify the adequacy of treatment integrity procedures were based on the thresholds set by Perepletchikova et al. (2007), which they describe as “consistent with the recommendations provided in the literature” (p. 834). However, the labels attached to each classification are arbitrary and do not necessarily reflect those used in current reporting guidelines for RCTs (e.g., Altman et al., 2001). Therefore, the extent to which our findings are replicated using other rating schemes remains to be investigated.

Ideally, future studies that compare treatments should adopt procedures to address each of the four domains of treatment integrity, as summarized by Perepletchikova et al. (2007). Treatment integrity should be established by the use of treatment manuals and therapist training. It should be assessed using direct observation and psychometrically sound measures. Measures of adherence should include not only prescribed tasks, but also
proscribed tasks, which allow for an evaluation of whether treatments are distinct from each other. Treatment integrity should be evaluated in a way that controls for bias (e.g., evaluate recordings of randomly chosen sessions). Numerical data on adherence and competence should be adequately reported for each treatment condition.

In this article, we argue that most of the comparative studies included in a recent meta-analysis examining the comparative efficacy of CBT and STPP did not adequately implement such procedures. Without adequate documentation of the steps taken to ensure and track treatment integrity, results of studies cannot be interpreted unambiguously. Studies should address issues of treatment integrity in order to more effectively answer the question about the comparative efficacy between CBT and other psychotherapies such as STPP.

REFERENCES


Received July 7, 2008; revised December 12, 2008; accepted December 12, 2008.