Journal of Consulting and Clinical Psychology
2007, Vol. 75, No. 6, 829–841

0022-006X/07/$12.00 DOI: 10.1037/0022-006X.75.6.829

Treatment Integrity in Psychotherapy Research: Analysis of the Studies and Examination of the Associated Factors

Francheska Perepletchikova, Teresa A. Treat, and Alan E. Kazdin
Yale University

Treatment integrity refers to the degree to which an intervention is delivered as intended. Two studies evaluated the adequacy of treatment integrity procedures (including establishing, assessing, evaluating, and reporting integrity; therapist treatment adherence; and therapist competence) implemented in psychotherapy research, as well as predictors of their implementation. Randomized controlled trials of psychosocial interventions published in 6 influential psychological and psychiatric journals were reviewed and coded for treatment integrity implementation. Results indicate that investigations that systematically addressed treatment integrity procedures are virtually absent in the literature. Treatment integrity was adequately addressed for only 3.50% of the evaluated psychosocial interventions. Journal of publication and treatment approach predicted integrity implementation. Skill-building treatments (e.g., cognitive–behavioral) as compared with non-skill-building interventions (e.g., psychodynamic, nondirective counseling) were implemented with higher attention to integrity procedures. Guidelines for implementation of treatment integrity procedures need to be reevaluated.

Keywords: treatment integrity, treatment fidelity, adherence, competence, treatment outcome

The main goals in treatment outcome research are specification of a treatment and evaluation of its feasibility and efficacy. Interpretations of treatment effects or the lack of treatment effects require some assurance that the treatment was carried out as it was designed (Kazdin, 2003). Treatment integrity (also known as treatment fidelity) refers to the extent to which the intervention was implemented as intended. Treatment integrity encompasses three aspects: (a) therapist treatment adherence, the degree to which the therapist utilizes prescribed procedures and avoids proscribed procedures; (b) therapist competence, the level of the therapist's skill and judgment; (c) and treatment differentiation, whether treatments differ from each other along critical dimensions (e.g., Waltz, Addis, Koerner, & Jacobson, 1993).

Therapist treatment adherence and treatment differentiation are closely related. A measure of therapist treatment adherence is sufficient for determination of whether treatments are in fact distinct (Waltz et al., 1993). When therapists adhere closely to the manual for each treatment (e.g., by implementing procedures prescribed for Treatment A and by avoiding procedures prescribed for Treatment B as well as other proscribed procedures), intervention purity is preserved. Manipulation checks on treatment delivery (i.e., assessment of treatment adherence) ensure that tasks pertaining to each treatment do not overlap.

Francheska Perepletchikova, Teresa A. Treat, and Alan E. Kazdin, Department of Psychology, Yale University.

This work was presented in part at the 18th Annual Convention of the Association for Psychological Science, New York, New York, May 28, 2006, and was supported in part by the Robert M. Leylan Dissertation Fellowship. We are very grateful to Susan Nolen-Hoeksema, Peter Salovey, and Douglas Mennin for their intellectual contributions and to Daniel J. Bauer for his help with statistical analyses.

Correspondence concerning this manuscript should be addressed to Francheska Perepletchikova at francheska.perepletchikova@yale.edu

The relationship between therapist treatment adherence and competence is less straightforward. Research examining the relationship between these two aspects has produced conflicting results, which range from no significant association (e.g., Paivio, Holowaty, & Hall, 2004) to high correlation (e.g., Barber et al., 2006; Shaw et al., 1999). Empirical examination of the association between therapist treatment adherence and competence may be challenging, due to the inherent conditionality between these two aspects: competence presupposes adherence, but adherence does not presuppose competence (McGlinchey & Dobson, 2003). Conceptual distinction, on the other hand, is more evident. Whereas adherence represents a quantitative aspect of treatment integrity (how frequently the therapist implements procedures prescribed by the manual and avoids those proscribed), competence is its qualitative aspect (how well prescribed procedures are implemented). Even if adherent, therapists can deliver treatment in an incompetent manner that threatens the validity of the interpretations about the obtained outcome. Failure to evaluate competence may result in an inability to establish which factors, treatment, or treatment provider resulted in the treatment effect or lack of effect. As noted by Nezu and Nezu (2005), the intervention may not equal the interventionist.

The breakdown in treatment integrity may pose threats to the experimental validity of a study and can have serious implications for inferences drawn about the relationship between treatment and outcome (e.g., Gresham, Donald, MacMillan, Beebe-Frankenberger, & Bocian, 2000; Kazdin, 2003; Moncher & Prinz, 1991). If a treatment was not executed as planned, it is not possible to establish which manipulation (intervention or alternative factors) resulted in a change on dependent measures, which would threaten the internal validity. Lack of treatment integrity can hinder attempts to replicate the study and to evaluate its external validity. Generality of the findings cannot be established without an exact description of what has actually been done to the depen-

dent variable. When an intervention is not provided as planned, the construct validity of the experiment is also compromised. Imprecision in intervention delivery can cause ambiguity in evaluating what the intervention was and why it produced the effect. Further, when treatment is not implemented as intended, unsystematic error may be introduced into the data, which compromises statistical conclusion validity. By increasing the within-group variability, such "noise" reduces the obtained effect size and statistical power and thus decreases the likelihood of detecting the effect.

This report consists of two studies. Study 1 evaluated the adequacy of treatment integrity procedures in the context of randomized controlled trials (RCTs) of psychotherapy published in influential psychiatric and psychological journals. Study 2 examined factors that were potential correlates of the implementation of integrity procedures, including treatment approach, corresponding author's educational background, the number of treatment comparisons, treatment characteristics, article type, and journal of publication.

Study 1: Analysis of Treatment Outcome Studies

Multiple recommendations have been provided in the literature on implementation of treatment integrity procedures (e.g., Carroll & Nuro, 2002; Gresham, 1997; Gresham et al., 2000; Schlosser, 2002; Waltz et al., 1993). These recommendations can be divided into four domains: establishing, assessing, evaluating, and reporting integrity. Establishing treatment integrity encompasses the operational definition of an intervention and the training and supervision of therapists. Treatment integrity depends on the completeness and clarity of the criteria that define the intervention (Kazdin, 2003). Detailing treatments in a manual reduces the variability in treatment implementation and enhances treatment integrity (e.g., Drozd & Goldfried, 1996). However, clear and unambiguous specification of the independent variable does not ensure that the manipulation will be implemented as planned without careful training of therapists. Training procedures can be roughly divided into indirect and direct categories (e.g., Sterling-Turner, Watson, Wildmon, Watkins, & Little, 2001). The indirect category includes didactic instructions about the intervention and written materials describing the rationale, scripts, and activities. The direct category includes opportunities for practice and involves procedures such as role-playing, rehearsal, feedback, and periodic booster sessions. A faithful rendition of the treatment is more likely with direct training procedures (e.g., Kratochwill, Elliott, & Busse, 1995; Sterling-Turner, Watson, & Moore, 2002). Therapists have to be supervised continuously to ensure accuracy of treatment implementation and to reduce therapeutic drift, which refers to gradual deviation from the treatment protocol (Kazdin, 2003).

Treatment integrity can be assessed via direct, indirect, and hybrid strategies. Direct observations can be conducted by trained staff present in the treatment setting, who view sessions through a one-way mirror, via monitors, or by videotaping. Indirect methods include therapist self-reports, debriefing clients on what was done during the treatment sessions, written homework assignments, and data collection sheets. Although these methods are less costly and laborious than are direct strategies, they are subject to distortion in self-representation, altered perception of the past, and poor recollection. Research that relies primarily on indirect evaluations of treatment integrity is likely to be weak in its ability to measure integrity accurately. As indirect measures of integrity offer immediate access to therapist adherence and to competence levels (Bergan & Kratochwill, 1990; Gresham, 1989), they can be used to supplement observational data and to adjust implementation (e.g., by directing therapist attention to omitted material and by encouraging the practice of inadequately executed procedures). Performance feedback may increase integrity when low levels are detected during treatment sessions (Codding, Feinberg, Dunn, & Pace, 2005).

Assessment of treatment integrity should encompass all three aspects involved in its specification: therapist treatment adherence, therapist competence, and treatment differentiation (Waltz et al., 1993). Therapist treatment adherence measures are sufficient for evaluation of treatment differentiation but only if they include proscribed procedures (i.e., procedures to avoid, as they may dilute intervention purity) as well as prescribed tasks. Therapist competence should not be assumed on the basis of experience and training but rather should be verified independently by measurement of how sensitively the treatment protocol is applied to individual clients. Data on the validity and reliability of integrity measures should be presented (see Perepletchikova & Kazdin, 2005 for discussion of validation methods).

Evaluation of treatment integrity encompasses procedures such as ensuring the accuracy of the representation of the obtained integrity data, training of raters, assessing interrater reliability, and controlling for measure reactivity. Accuracy of the representation of integrity levels depends upon the collection of data across treatment phases, therapists, situations, sessions, and cases. For example, some treatment phases (e.g., assessment of the pathology) may be simpler than are others (e.g., training of skills). Higher integrity ratings may be achieved when data are collected primarily during the administration of more straightforward tasks. Rater competence requires rigorous training in all of the major and minor treatment components, including subtle aspects of the treatment and the treatment manual. Thus, raters who are skilled in the delivery of the treatment being evaluated seem to be the most suitable for integrity rating. Regardless of who performs the ratings, interrater reliability checks are important for ensuring adequate assessment of integrity. Additionally, as any assessment may be reactive, reactivity should be assessed and controlled. Therapists' self-reports can be biased and distorted by self-interest. Observations can alter performance of the therapist and may result in higher adherence to specified procedures during the observed sessions. To ameliorate reactivity, staff can perform "spot checks" of treatment implementation on a variable time schedule, where therapists are interviewed by staff members at random times and without notification. Also, reactivity may be lower when all of the sessions are videotaped or observed, which may preclude the fluctuation in integrity due to the presence or absence of an observer.

Treatment integrity should be reported in terms of overall integrity, component integrity, and session integrity (Gresham, 1997; Schlosser, 2002). Overall integrity reflects the integrity of treatment components across sessions. Component integrity refers to the integrity of implementing each treatment component across sessions. Session integrity refers to the integrity of all treatment components within each session. Although overall integrity may be high, treatment may be implemented with low adherence and

competence due to poor component integrity and poor session integrity. For example, therapist performance may vary as a function of client difficulty (e.g., Patterson & Chamberlain, 1994), and such variability may result in inconsistent treatment delivery within sessions. Although all treatment components may be implemented across sessions, session integrity may be low.

Numerical data on adherence and competence should adequately describe the level of treatment integrity. Reporting that utilization of treatment components was significantly higher for one intervention than for another does not indicate absolute adherence levels (e.g., 50% integrity may be significantly higher than 20% integrity, but neither represents adequate integrity levels). Further, integrity sometimes is evaluated by asking raters to classify videotapes of therapy sessions by the employed treatment modality (e.g., which tape belongs to cognitive vs. interpersonal therapy). A tape may be correctly classified because the number of components within a session was higher for one treatment than for the other. However, this does not demonstrate that all of the prescribed components were utilized during a session or that proscribed interventions were not delivered. Only absolute values of therapist treatment adherence and competence levels accurately denote treatment integrity levels.

Recommendations discussed previously address how to establish, assess, evaluate, and report treatment integrity. Prior reviews have examined the implementation of some of the recommended procedures, including using treatment manuals, training and supervising therapists, and reporting numerical values of treatment integrity (e.g., Borrelli et al., 2005; Gresham et al., 2000; Gresham, Gansle, & Noell, 1993). These reviews focused primarily on whether integrity procedures were addressed in the literature. Such evaluation of treatment integrity offers limited insight into the quality of the employed procedures. If procedures were not executed adequately, the examination of whether they were performed may be misleading. For example, previous reviews indicated that only 6%–27% of studies reported and assessed integrity (e.g., Borrelli et al., 2005; Gresham, Gansle, Noell, Cohen, & Rosenblum, 1993; Rogers Wiese, 1992). However, assessing integrity using nonvalidated measures or reporting that integrity was monitored without providing quantitative information does not constitute appropriate implementation of integrity procedures. Evaluation of how treatment integrity procedures were implemented (e.g., therapist training strategies, direct vs. indirect assessment methods, validity and reliability of utilized measures) addresses the question of the adequacy of the integrity procedures.

Furthermore, prior reviews provided a fragmented examination of treatment integrity. They evaluated implementation of each procedure separately without considering the overall extent to which integrity was addressed in research (i.e., the percent of interventions tested with adequate attention to integrity to allow unambiguous interpretation of the obtained results). Such evaluation is imperative, as it informs judgments of the overall quality of psychotherapy research.

We had three objectives in Study 1. First, we examined the overall extent to which treatment integrity was addressed in RCTs published in influential psychiatric and psychological journals. We restricted the scope of the current review to these journals in order to provide an evaluation of treatment integrity implementation under the best possible circumstances. That is, these journals can be considered the "gold standard" for reporting findings from treatment outcome research, as they maintain stringent criteria related to documentation of research methodology. Second, we evaluated the adequacy of treatment integrity procedures implemented in four domains of integrity (establishing, assessing, evaluating, and reporting integrity). Finally, we evaluated the adequacy of integrity procedures in its two main aspects, therapist treatment adherence and therapist competence.

Method

Literature search procedures. To identify journals for review, we searched PsycINFO with a combination of 25 psychotherapy-related key terms (e.g., *treatment*, *therapy, intervention*) and 75 descriptive terms (e.g., *psychosocial, cognitive, dynamic*); the complete list of terms is available on the treatment integrity website or from the corresponding author.[1] The resulting search was limited by the terms *peer reviewed journal, human, English language,* and *years 2000–2004*. MEDLINE was not consulted because the present study focused on psychosocial interventions, and the journals of interest (see journal selection criteria below) were indexed in PsycINFO.

Criteria for journal selection were as follows: (a) frequent publication of treatment outcome research (≥100 articles) between the years 2000 and 2004 and (b) consistent listing in the top 10 influential journals in psychiatry or clinical psychology (by impact factor) in the years 2000 to 2004 by *Thomson ISI Journal Citation Reports*.[2] The 6 journals that met the above criteria were the *Archives of General Psychiatry* (*AGP;* No. 17 by number of occurrences of treatment outcome studies, 192 studies); the *American Journal of Psychiatry* (*AJP;* No. 3, 295); the *British Journal of Psychiatry* (*BJP;* No. 6, 196); the *Journal of the American Academy of Child and Adolescent Psychiatry* (*JAACAP;* No. 8, 170); the *Journal of Consulting and Clinical Psychology* (*JCCP;* No. 7, 192); and the *Journal of Clinical Psychiatry* (*JCP;* No. 1, 479). *AGP, AJP, JCP,* and *BJP* were consistently listed among the first 5 psychiatric journals for the years 2000–2004 by impact factor. *JCCP* was consistently listed among the first 3 clinical psychology journals for the years 2000–2004. *JAACAP* was consistently listed among the first 9 psychiatric journals.

After journals were identified, all of the articles within each journal were examined and hand selected for inclusion in the

[1] The supporting materials can be obtained from the treatment integrity website, www.treatmentintegrity.com (the title of this manuscript serves as a link), or from the corresponding author. Available materials include the list of terms for the literature search procedures, the list of the evaluated studies, two measures (ITIPS and AVC), the ITIPS rater manual, the scoring procedures for the treatment approach categories, the Checklist of Treatment Integrity Procedures, and additional tables.

[2] Psychiatric and clinical psychology journals were listed in separate subject disciplines in the Thomson ISI report (2002). An impact factor represents a relationship between a journal and an average of similar journals that cover a subject discipline. It is a measure of the frequency with which the "average article" in a journal has been cited in a particular year and is calculated by dividing the citation impact for a journal in a particular field by citation impact for the field as a whole, worldwide. The mean impact factors for the selected journals for the years 2000–2004 are as follows: *AGP* ($M = 11.42$, $SD = .58$), *AJP* ($M = 6.94$, $SD = .46$), *BJP* ($M = 4.39$, $SD = .27$), *JAACAP* ($M = 3.55$, $SD = .23$), *JCCP* ($M = 3.81$, $SD = .48$), and *JCP* ($M = 4.66$, $SD = .26$).

sample. RCTs were included in the sample if they (a) assessed the effect of a psychosocial intervention on a set of dependent measures (some studies compared several psychosocial treatments or juxtaposed psychosocial and pharmacological treatments); (b) included a comparison of psychosocial intervention(s) to a control group (wait list, no treatment, placebo, treatment as usual, or other procedures intended to be a control condition); (c) utilized a prospective design and random assignment of subjects to conditions; (d) used participants selected for having psychological problems; and (e) included posttreatment assessment of therapeutic change. The list of the evaluated studies can be obtained from the treatment integrity website or from the corresponding author.

Articles were excluded from the sample if they (a) had a primary purpose other than the evaluation of the effects of a psychosocial intervention on dependent measure(s), including examination of the predictors of study outcome, mediators or moderators of therapeutic process, risk factors, cost effectiveness of the intervention, barriers to treatment implementation, characteristics of sample and treatment setting, and follow-up studies; (b) evaluated active interventions that were not delivered by treatment agents (e.g., bibliotherapy, computerized or mail-based therapies, self-help therapies); (c) made comparisons between highly standardized and low standardized treatments; and (d) evaluated only pharmacological interventions. Overall, 147 articles, evaluating 202 treatments, were identified. The breakdown by journals was as follows: *AGP* (22 articles, 25 treatments), *AJP* (9 articles, 9 treatments), *BJP* (19 articles, 29 treatments), *JAACAP* (16 articles, 19 treatments), *JCCP* (75 articles, 113 treatments), and *JCP* (6 articles, 7 treatments).

Measure. We developed the Implementation of Treatment Integrity Procedures Scale (ITIPS) to evaluate the extent to which RCTs addressed the four domains of treatment integrity, including establishing, assessing, evaluating, and reporting integrity, as well as its two main aspects, therapist treatment adherence and therapist competence; the ITIPS and the scoring manual can be obtained from the treatment integrity website or from the corresponding author. The ITIPS consists of 22 items, each rated on a 4-point scale. Total scores range from 22 to 88. Higher scores indicate more adequate implementation of integrity procedures (e.g., "Training strategies of therapists," where 1 = *not trained,* 2 = *authors mentioned that therapists were trained but no other information was provided,* 3 = *used indirect strategies,* and 4 = *used direct strategies*).

The establishing treatment integrity domain (6 items) refers to how researchers conceptualize integrity (e.g., in terms of adherence and/or competence), as well as the extent to which they provide a detailed treatment manual to therapists, train them, and supervise them. The assessing treatment integrity domain (7 items) refers to the assessment of treatment integrity via direct, indirect, or hybrid strategies; measurement of therapist treatment adherence as well as competence; and employment of integrity measures with good psychometric properties (i.e., validity and reliability). The evaluating treatment integrity domain (5 items) refers to procedures such as ensuring the accuracy of the representation of the obtained integrity data, training of raters, assessing interrater reliability, and controlling for measure reactivity. The reporting treatment integrity domain (4 items) refers to procedures such as reporting numerical data; reporting overall, component, and session integrity; and reporting the implementation of various integrity procedures. Therapist treatment adherence and therapist competence aspects of integrity (6 items each) encompass how the terms were defined, assessed, evaluated, and reported.

Training of raters. We rated articles on the implementation of treatment integrity procedures using a specific manual that was developed for this study. The manual contains general information about treatment integrity and scoring procedures for each item on the ITIPS and includes specific examples from the literature. The principal investigator (F. P.) trained two undergraduate students (both female, 18 and 20 years of age, one African American, one Hispanic American) as raters. Following Gresham et al.'s design (1993), a series of four, 1-hr sessions spread over 4 weeks was implemented in the training of raters; no articles used for training purposes were included in the sample. After training, raters were supervised weekly by the principal investigator to prevent rater drift and to clarify any questions pertaining to article rating.

The two raters scored articles independently. Inconsistencies in scoring were resolved by consensus between raters. We calculated interrater agreement using the *T* index, which allows evaluation of rater agreement along ordinal scales and accounts for chance agreement (Tinsley & Weiss, 2000).[3] Agreement was defined as identical scores on an item on a 4-point scale. The *T* index for the preconsensus ratings was .73. The *T* index of postconsensus ratings with the ratings of the principal investigator was .89 (rater consensus scores were used for the analyses). When the magnitude of the T index of interrater agreement on the ITIPS is examined, the nature of the evaluated information should be considered. Report of treatment integrity procedures is very low and is not standardized. High variability in reporting strategies, coupled with low rates of implementing integrity procedures, required the rater to sift through an entire article (not just the Method and Results sections) for any relevant information (frequently outlined in just one or two words) and to make inferences on the adequacy of the integrity procedures from barely sufficient data. Such a task may be particularly susceptible to error by omission. Therefore, consensus ratings were obtained for each treatment.

Data evaluation procedures. We used ITIPS to evaluate the degree to which treatment integrity procedures were addressed in the articles. When authors referred to outside sources for further information regarding the implemented integrity procedures (e.g., description of a manual, validity of integrity measures), we consulted these sources to make informed ratings; references were consulted for 36 treatments. Authors of the articles were not contacted for information, and none of the articles specified availability of the additional information on integrity from the authors.

The adequacy of integrity procedures was examined by evaluation of scores on the ITIPS across 202 treatments for total integrity, its four domains, and two aspects. On each item, a score of 1 or 2 was assumed to reflect inadequate implementation of

[3] The *T* index is calculated using the following equation, $T = (Na - N\rho c) / (N - N\rho c)$, where Na = the number of agreements, N = the number of agreements plus the number of disagreements, and ρc = the probability of chance agreement on an item. Tinsley and Weiss (2000) provide the probability of chance agreement when agreement is defined as a 0-point, 1-point, or 2-point discrepancy. Positive values of *T* indicate that rater agreement is greater than chance, negative values indicate lower than chance agreement, and *T* is zero when rater agreement is equal to chance agreement.

Table 1
Percentage Adequacy of the Implementation of Treatment Integrity Procedures

Variable	Total treatment integrity			Establishing treatment integrity			Assessing treatment integrity			Evaluating treatment integrity			Reporting treatment integrity		
	IA	AA	AD	IA	AA	AD	IA	AA	AD	IA	AA	AD	ID	AA	AD
Overall (%)	60.40	36.10	3.50	36.60	47.50	15.80	77.20	19.30	3.50	87.60	10.40	2.00	58.90	34.70	6.40
Mean score	31.03	52.25	69.00	8.47	15.81	20.18	9.60	17.00	22.00	5.93	12.43	16.75	5.72	10.57	13.77
(*SD*)	(7.54)	(5.48)	(1.15)	(1.95)	(1.64)	(1.42)	(2.86)	(1.86)	(0.00)	(1.35)	(1.25)	(0.50)	(1.39)	(0.02)	(0.44)
Range	22–44	45–66	67–70	6–12	13–18	19–24	7–14	15–20	22	5–10	11–15	17	4–8	9–12	13–14
AGP (%)	60.00	40.00	0.00	32.00	48.00	20.00	76.00	20.00	4.00	84.00	16.00	0.00	68.00	24.00	8.00
AJP (%)	77.80	22.20	0.00	66.70	22.20	11.10	100.00	0.00	0.00	100.00	0.00	0.00	88.90	11.10	0.00
BJP (%)	86.20	13.80	0.00	55.20	44.80	0.00	93.10	3.40	3.40	100.00	0.00	0.00	69.00	31.00	0.00
JAACAP (%)	73.70	21.10	5.30	42.10	42.10	15.80	94.70	5.30	0.00	78.90	15.80	5.30	57.90	36.80	5.30
JCCP (%)	48.70	46.00	5.30	26.50	53.10	20.40	67.30	29.20	3.50	85.00	12.40	2.70	49.60	41.60	8.80
JCP (%)	85.70	14.30	0.00	85.70	14.30	0.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
M across journals	72.02	26.23	1.77	51.37	37.41	11.22	88.52	9.65	1.82	91.32	7.37	1.33	73.23	24.08	3.68
(*SD*)	(14.92)	(13.56)	(2.75)	(22.40)	(15.49)	(9.32)	(13.64)	(12.11)	(2.00)	(9.73)	(8.17)	(2.22)	(18.95)	(15.91)	(4.20)

Note. IA = inadequate implementation; AA = approaching adequacy; AD = adequate; *AGP = Archives of General Psychiatry; AJP = American Journal of Psychiatry; BJP = British Journal of Psychiatry; JAACAP = Journal of the American Academy of Child and Adolescent Psychiatry; JCCP = Journal of Consulting and Clinical Psychology; JCP = Journal of Clinical Psychiatry.*

integrity procedures; a score of 3 indicated that implementation approached adequacy; and a score of 4 designated adequate implementation of integrity procedures. Because there were 22 items on the ITIPS, studies were classified as implementing integrity procedures (a) inadequately if the study's total score ranged between 22 and 44; (b) in a manner approaching adequacy if the total score ranged between 45 and 66; and (c) adequately if the total score exceeded 66. This strategy was also utilized for evaluation of the adequacy of the treatment integrity procedures for the four domains and the two aspects of integrity (see Tables 1 and 2 for ranges). The percentage of treatments implementing integrity procedures within each range of scores was calculated.

Results

Internal consistency of the ITIPS. To evaluate how well the items on the ITIPS related to the overall score, we computed individual item-remainder correlations. In each case, the score on an item was correlated with the total score, after that item was removed from the total. The item-remainder score correlations of the 22 items were positive and ranged from .42 to .90 ($M = .66$, $SD = .31$, $p < .001$). Moreover, the correlations between scores on the four domains of treatment integrity were positive and ranged from .53 to .93 ($M = .77$, $SD = .34$, $p < .001$). The correlation of therapist treatment adherence and therapist competence scores was .35 ($p < .001$). The 22-item ITIPS ($M = 40.01$, $SD = 13.33$, range 22.00–70.00) demonstrated good internal consistency. Cronbach's alpha and the Spearman–Brown coefficient were .94 and .89, respectively. For the four integrity domains and two integrity aspects, Cronbach's alpha ranged from .76 to .93 ($M = .84$, $SD = .07$), and the Spearman–Brown coefficient ranged from .74 to .97 ($M = .86$, $SD = .09$).

Adequacy of implementing treatment integrity procedures across treatments. As can be seen in Table 1, the ratings for the implementation of treatment integrity procedures (total score on ITIPS) were 60.40% inadequate, 36.10% approaching adequacy, and 3.50% adequate. Treatment integrity was established 36.60% inadequately, 47.50% with approaching adequacy, and 15.80% adequately. Treatment integrity was assessed 77.20% inadequately, 19.30% with approaching adequacy, and 3.50% adequately. Treatment integrity was evaluated 87.60% inadequately, 10.40% with approaching adequacy, and 2.00% adequately. Across interventions, treatment integrity levels were reported 58.90% inadequately, 34.70% with approaching adequacy, and 6.40% adequately.

Table 2 documents that treatment adherence procedures were implemented across treatments 52.00% inadequately, 39.10% with approaching adequacy, and 8.90% adequately. Therapist competence procedures were implemented 87.10% inadequately, 11.40% with approaching adequacy, and 1.50% adequately. Across therapies, a treatment manual was not mentioned 14.40% of the time and was only mentioned, without provision of any details pertaining to the treatment protocol, 3.00% of the time. The ratings for "manual is general" and "manual is specific" were 17.30% and 65.30%, respectively.

Discussion

The results indicate that investigations that systematically implement treatment integrity procedures are extremely rare in the

Table 2
Implementation of Therapist Treatment Adherence and Therapist Competence Procedures and Utilization of Treatment Manual

Variable	Therapist treatment adherence procedures			Therapist competence procedures			Utilization of treatment manual			
	IA	AA	AD	IA	AA	AD	Manual not mentioned	Manual only mentioned	Manual is general	Manual is specific
Overall (%)	52.00	39.10	8.90	87.10	11.40	1.50	14.40	3.00	17.30	65.30
Mean score	7.20	15.24	19.44	6.26	16.17	19.00				
(*SD*)	(2.02)	(1.78)	(0.50)	(1.10)	(1.87)	(0.00)				
Range	6–12	13–18	19–20	6–12	13–18	19				
AGP (%)	56.00	40.00	4.00	84.00	16.00	0.00	20.00	0.00	28.00	52.00
AJP (%)	89.00	11.00	0.00	100	0.00	0.00	22.20	0.00	33.30	44.40
BJP (%)	65.50	34.50	0.00	93.10	6.90	0.00	34.50	0.00	0.00	65.50
JAACAP (%)	57.90	36.80	5.30	94.70	5.30	0.00	0.00	15.80	10.50	73.70
JCCP (%)	41.60	44.20	14.20	84.10	13.30	2.70	6.20	2.70	20.40	70.80
JCP (%)	85.70	14.30	0.00	86.70	14.30	0.00	71.40	0.00	0.00	28.60
M across journals	65.95	30.13	3.92	90.43	9.30	0.45	25.72	3.08	15.37	55.83
(*SD*)	(18.32)	(13.97)	(5.54)	(6.52)	(6.23)	(1.10)	(25.50)	(6.32)	(14.16)	(17.46)

Note. IA = inadequate implementation; AA = approaching adequacy; AD = adequate; *AGP* = *Archives of General Psychiatry*; *AJP* = *American Journal of Psychiatry*; *BJP* = *British Journal of Psychiatry*; *JAACAP* = *Journal of the American Academy of Child and Adolescent Psychiatry*; *JCCP* = *Journal of Consulting and Clinical Psychology*; *JCP* = *Journal of Clinical Psychiatry*.

literature. Although it is generally recognized that a fair test of an intervention is impossible without determination of whether it was conducted as designed, integrity was adequately addressed for only 3.50% of the treatments. This may mean that observed changes on the dependent measures could be unambiguously interpreted for only 3.50% of the evaluated treatments. Further, separate evaluations of the four domains of treatment integrity (establishing, assessing, evaluating, and reporting integrity) indicated insufficient implementation of the procedures in all domains. Similarly, both aspects of treatment integrity (adherence and competence) were inadequately addressed in the literature.

Treatment integrity procedures "approached adequacy" for 36.10% of treatments. However, the approaching adequacy category was used only for descriptive purposes, as it means just as little as results that are "approaching significance." For example, the adequacy of providing a treatment manual without training therapists on its implementation may be questioned, as this approach does not ascertain that the intervention is accurately applied.

It can be argued that the obtained outcome is a function of the high adequacy criteria and that a lower threshold would be a more reasonable standard. However, the criteria utilized in the current study are consistent with the recommendations provided in the literature and reflect standards for assessment of dependent variables (e.g., training of raters, valid and reliable measures, interrater reliability). Measures of outcome receive far more attention than does treatment integrity. Peterson, Homer, and Wonderlich (1982) described this phenomenon as a "curious double standard" (p. 478), in which operational definitions and measures of reliability are detailed when behaviors serve as dependent variables and are virtually ignored when behaviors serve as independent variables. Setting a threshold for implementation of treatment integrity procedures lower than that established for measures of outcome may help perpetuate the double standard, instead of pointing to the discrepancy in the way experimental variables are addressed.

Further, Study 1 evaluated the utilization of treatment manuals. Although provision of a specific manual is an integral part of the methodology of RCTs (e.g., Chambless & Hollon, 1998; Westen, Novotny, & Thompson-Brenner, 2004), use of a specific protocol was reported in only 65.30% of treatments. Not all therapies can be specified to the degree that is required for empirical testing. For some treatments (e.g., humanistic, client-centered), creative responding and improvisation in the moment are valued as key ingredients of therapeutic process and specific protocols are viewed as counterproductive (Bohart, O'Hara, & Leitner, 1998). The need for well-tested manuals is also questioned, because therapists do not seem to use them in clinical practice (e.g., Addis, Wade, & Hatgis, 1999; Wilson, 1998). Further, with clinically complex cases (e.g., comorbidity), therapists may need to incorporate interventions not outlined in a manual to address specific concerns (Henry, 1998). However, recent evidence suggests that the degrees of therapist flexibility and treatment tailoring do not predict favorable outcomes (Kendall & Chu, 2000). Whether or not the arguments against the use of treatment manuals are valid, current methods for empirical testing rely on standardization, which minimizes within-group variability, controls confounding variables, reduces ambiguity when interpreting outcomes, and allows for replication of results.

The lack of studies that adequately address treatment integrity undermines our confidence in psychotherapy research. It is possible that the pattern of results in the current study may reflect incomplete reporting of the procedures utilized by the researchers or deletion of the treatment integrity information by the authors or journal editors before a study is published. However, as noted by Johnston and Pennypacker (1980), incomplete reporting of procedures puts readers in the position of giving authors the benefit of the doubt or of refusing to accept provided inferences. So far, as Study 1 suggests, consumers of research are expected to have fairly unbridled confidence in therapists' ability to deliver treatments as intended without manipulation checks. Once understood, the reasons why inadequate attention is paid to treatment integrity may provide clues pertaining to possible interventions to improve the situation. Testing predictors of integrity implementation may offer insights into what contributes to the problem.

Study 2: Predictors of Integrity Implementation

Predictors of treatment integrity implementation may encompass multiple factors, including treatment approach, educational background of the corresponding author, number of active interventions tested in a study, and type of experimental report. The degree to which treatment integrity is addressed may be associated with the nature of the therapeutic approach tested in a study. The more specific and concrete the treatment, the easier it is to operationalize the intervention and to monitor its implementation. The complexity of therapy is inversely related to the level of treatment integrity (Gresham et al., 2000). Treatment complexity refers to the number and specificity of treatment components. That is, the higher the number and the more abstract the components, the more complex the intervention. Interventions that are complex might be specifically at risk for procedural degradation, because of the increased difficulty of establishing, evaluating, and maintaining integrity.

Skill-building approaches (e.g., cognitive–behavioral interventions) may be viewed as less complex than are process-oriented (e.g., psychodynamic, humanistic therapies) and supportive–educational approaches (e.g., nondirective counseling, psychoeducation). Skill-building therapies are more straightforward and simple, as they are intended to alleviate a disorder by solving presented problems through specific strategies and techniques that can be organized in terms of a treatment protocol. Process-oriented approaches, on the other hand, focus more on acquisition of insight and exploration of underlying dynamics, rather than on learning specific techniques. Supportive–educational interventions may incorporate several therapeutic modalities, and the content of the material presented in a given session may be based primarily on individuals' needs, rather than on a specific treatment plan. In such therapies, flexible, spontaneous responding in the moment is valued (Bohart et al., 1998). The skillful improvisation required in non-skill-building therapies may render specific operational definition of an intervention extremely difficult and even counterproductive. For example, manualization of psychodynamic therapy may actually lead to deterioration in certain aspects of therapists' interpersonal skills (Binder, 1993). Manuals developed for non-skill-building approaches are usually general. When a detailed protocol is not provided, therapist performance can vary widely from occasion to occasion. Thus, unsystematic and random variation is introduced into the delivery of the treatment, and the task of establishing and monitoring integrity becomes increasingly difficult. Thus, non-skill-building approaches may address treatment integrity procedures to a lesser degree.

The extent to which treatment integrity procedures are implemented may also relate to the educational background of the corresponding author of the articles. The corresponding author is most often the senior investigator for the study and oversees all aspects of the project. Authors with research degrees may be more likely to have received training in research methodology procedures than are authors with medical or clinical degrees. Therefore, authors with research degrees (e.g., PhD), as compared with nonresearch degrees (e.g., MD), may implement treatment integrity procedures at a higher level.

The number of active interventions tested in a study may also contribute to integrity implementation. When several treatments are compared, it is important to establish treatment differentiation, as it ensures that treatments were not diluted by incorporating components from other interventions. The diluted interventions may fail to show differential effects on dependent measures, because their components are no longer distinct. Adequately addressing therapist treatment adherence as a part of treatment integrity procedures ensures treatment differentiation (Waltz et al., 1993). Therefore, when two or more active treatments are compared, treatment integrity may be addressed to a greater extent than when only one intervention is evaluated. Further, the type of experimental report may affect treatment integrity. Full reports, as compared with brief reports, are usually less constrained in the amount of space that can be devoted to the description of employed methods. Therefore, it may be predicted that treatment integrity procedures will be addressed to a higher degree in full reports than in brief reports.

On the basis of previous information, four hypotheses were formulated. First, studies that evaluate skill-building approaches, relative to studies that evaluate non-skill-building approaches, will address treatment integrity procedures to a greater degree. Second, authors with research degrees will implement more integrity procedures than will authors with nonresearch degrees. Third, studies that evaluate two or more psychosocial interventions will implement more treatment integrity procedures than will studies that examine one therapy. Fourth, full reports will address treatment integrity to a higher degree than will brief reports.

Additionally, five questions for which specific hypotheses were not formulated were examined for exploratory purposes. First, does the journal of publication relate to treatment integrity implementation? Second, does the number of years since the corresponding author received his or her highest degree relate to the extent to which integrity procedures are implemented? Third, does the extent to which treatment integrity is addressed relate to the country in which the corresponding author's degree was received? Fourth, does treatment type (individual versus nonindividual therapy) relate to integrity implementation? Finally, does treatment duration affect the extent to which integrity is addressed?

Method

Measures. We developed the Associated Variables Checklist (AVC) for this study to examine predictors of treatment integrity implementation; the AVC and the scoring procedures can be

obtained from the treatment integrity website or from the corresponding author. Examined predictors include the treatment approach, corresponding author's educational background (highest degree, years from degree, and country where degree was received), number of treatment comparisons, treatment characteristics (type and duration), article type, and journal of publication. Treatment approach was classified in one of four categories: process oriented (e.g., psychodynamic, existential), supportive–educational (e.g., nondirective counseling, motivational interviewing), skill training (e.g., cognitive–behavioral interventions), and other approaches (e.g., vocational rehabilitation). The "other approaches" category ($n = 5$) was of no conceptual interest, and treatments belonging to this category were removed from the analyses. Thus, of the 202 therapies identified for Study 1, present analyses included 197 treatments. A dichotomous treatment approach variable was created, where 0 = "non-skill-building approaches" (process-oriented and supportive–educational) and 1 = "skill-building approaches."

The corresponding author's highest degree was rated on a 2-point scale, where 0 = "nonresearch degree" (in North America: MD, PsyD, MSW; in England: MB, BS) and 1 = "research degree" (in North America: PhD; in England: DM and MD). Authors who had both nonresearch and research degrees were coded as having a research degree. Years from degree was calculated by subtracting the year when the degree was received from the year of publication. Country of degree was rated on a 2-point scale, where 0 = "outside of the United States" and 1 = "inside the United States."

Number of treatment comparisons was rated on a 2-point scale, where 0 = "one active treatment compared with control condition" and 1 = "two or more comparisons between active treatments." Treatment type was divided into two categories: 0 = "nonindividual therapy" (e.g., group, couples, family, and dyads) and 1 = "individual therapy." Treatment duration represented the total number of sessions per treatment. A session of 45–90 min was counted as one session; when sessions extended for more than 90 min, the total number of sessions was multiplied by the number of hours per session (e.g., five sessions × 3 hr each). Article type was rated on a 2-point scale, where 0 = "brief report" and 1 = "full report." We gave each journal (i.e., *AGP*, *AJP*, *BJP*, *JAACAP, JCCP*, and *JCP*) a number from 1 to 6 to code journal of publication.

Data collection procedures. The two undergraduate students who served as raters for Study 1 collected and coded data from each study for the following variables: number of treatment comparisons per study, treatment characteristics (type and duration), article type, and journal of publication. Information on corresponding author's highest degree, years from degree, and country where this degree was received was collected by emailing or calling corresponding authors, consulting faculty websites of universities, and searching through the online Biography Resource Center. Obtaining and coding this information did not require subjective evaluation or interpretation on the part of the rater. Coding treatment approaches evaluated in the studies, on the other hand, required subjective judgments; therefore, two graduate students were used as raters on this variable in place of undergraduate students.

The principal investigator trained two advanced graduate students in clinical psychology (one male, one female, 24 and 25 years of age, one Hispanic American, one European American) to perform the ratings of treatment approaches. Before the training session, raters were asked to read the scoring procedures that described criteria for rating each treatment approach category. During the 1-hr training session, rating criteria were discussed; raters independently scored 20 treatments (that were not part of the study) and arrived at consensus ratings for each treatment. The two trained raters independently scored each treatment evaluated in the study. Any inconsistencies in scoring were resolved by consensus between raters. Interrater agreement was defined as identical scores on an item on a 4-point scale. The *T* index for the preconsensus ratings was .92. The *T* index of the postconsensus ratings with the ratings of the principal investigator was .91. Rater consensus scores were used for the analyses.

Data evaluation procedures. Prior to analysis, missing values and adherence of variable distributions to the assumptions of multivariate analyses were checked with SPSS 12.0. There were no missing values in the data set, and variables were normally distributed, except for treatment duration. To reduce the extreme skewness of the treatment duration variable (skewness = 1.74, $SE = .17$), we applied a square-root transformation, which made the distribution approximately normal.[4]

The analyses were performed on 197 treatments. Five interventions that belonged to the "other approaches" treatment category were removed from the original sample of 202 therapies for all analyses, as this category was not of conceptual interest.

We employed hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992) to examine the influence of predictor variables on treatment integrity implementation. HLM was used due to the nested or clustered data structure, as treatments were nested within studies. In the present data set, the number of observations at the study level ranged from one to three (i.e., there were one to three treatments per study). Utilization of ordinary-least-squares regression procedures under these conditions violates the independence assumption and deflates the estimated standard errors. To examine the degree of clustering in the data, we computed the intraclass correlation (ICC) on treatment integrity, as the outcome variable. Treatment integrity was measured by the total score on the ITIPS ($M = 39.75$, $SD = 13.10$, range 22.00–70.00). The ICC in the present study was .96, which indicated that 96% of the total variance in the treatment integrity scores was between study, whereas about 4% of the variance was within study (i.e., between treatments within a single study). An indirect test of the significance of the ICC rejected the null hypothesis that between-study variability was zero, $\chi^2(1, N = 144) = 5{,}002.34$, $p < .001$ (Heck & Thomas, 2000).

Results

Treatment approach. We predicted that treatment integrity would be addressed to a higher degree when skill-building approaches ($n = 145$) were evaluated in a study as compared with non-skill-building approaches ($n = 52$). To test this effect, we specified an HLM equation in which treatment approach was the predictor and treatment integrity was the outcome. Results were

[4] Our data analyses in HLM without transformation of the treatment duration variable, using robust standard errors, did not affect the pattern of statistical significance for the obtained results for any of the predictors.

consistent with our prediction; treatment integrity procedures were implemented to a greater extent when skill-building treatments were evaluated, as compared with non-skill-building therapies, $t(195) = 2.16$, $p < .05$, $r^2 = .02$.[5]

Corresponding author's educational background. To examine the relation between the corresponding author's educational background and treatment integrity, we specified a separate HLM equation for each of the educational background variables (highest degree, years from degree, and the country of degree). Each variable served as a predictor, and treatment integrity was the outcome. Author's degree, $t(142) = 3.03$, $p < .01$, $r^2 = .05$, and country of degree, $t(142) = 2.42$, $p < .05$, $r^2 = .03$, were related to treatment integrity. Results indicated that authors with research degrees ($n = 110$) implemented more treatment integrity procedures than did authors with nonresearch degrees ($n = 34$). Further, authors who received their degrees in the United States ($n = 79$) were more likely to address treatment integrity than the authors who received their degrees outside of the United States ($n = 65$). Years from degree ($M = 16.98$, $SD = 10.70$, range 0–46) did not predict treatment integrity implementation, $t(142) < 1.00$, *ns.*

Journal of publication. Journal of publication was a nominal-scale variable with five categories. It was converted into four dichotomous dummy variables for analyses, as dichotomous variables can be appropriately tested by linear methods, such as HLM (Tabachnick & Fidell, 2001). The five dummy variables were *AGP* ($n = 21$), *AJP* ($n = 9$), *BJP* ($n = 19$), *JAACAP* ($n = 19$), and *JCP* ($n = 6$). *JCCP* ($n = 73$) was used as the reference category. An HLM equation was specified, in which journals of publication were the predictors and treatment integrity was the outcome. The results indicated that treatment integrity was predicted by the journal of publication, $\chi^2(5, N = 144) = 4{,}420.04$, $p < .001$, $r^2 = .07$. Integrity procedures were implemented to a greater degree in *JCCP*, relative to *AJP,* $t(138) = -3.37$, $p < .001$; *BJP,* $t(138) = -2.79$, $p < .01$; and *JCP,* $t(138) = -3.37$, $p < .001$. There was no difference between *JCCP* and *AGP,* $t(138) < 1.00$, *ns.* The difference between *JCCP* and *JAACAP* was also not significant, $t(138) = -1.34$.

Nonsignificant results. Studies evaluating two or more active psychosocial therapies ($n = 45$), as compared with studies examining one psychosocial therapy ($n = 99$), did not address treatment integrity procedures to a greater extent, $t(142) < 1.00$, *ns.* Individual therapies ($n = 97$), as compared with nonindividual therapies ($n = 100$), did not predict the degree of treatment integrity implementation, $t(195) < 1.00$, *ns*. Similarly, treatment duration ($M = 13.97$, $SD = 10.64$, range 1–56) was not related to treatment integrity implementation, $t(195) < 1.00$, *ns.* Full reports ($n = 129$) did not address treatment integrity procedures to a higher degree than did brief reports ($n = 15$), $t(142) = 1.56$, *ns.*

Combined effect of all significant predictors on the treatment integrity levels. To evaluate the combined effect of the identified significant predictors on treatment integrity implementation, we specified an HLM equation in which all significant predictors were entered simultaneously. Due to the pattern of results, journal of publication was converted into a dichotomous variable, where *AJP*, *BJP,* and *JCP* represented one category (coded as 0), and *AGP, JAACAP,* and *JCCP* represented another category (coded as 1).

The results indicated that treatment approach and journal of publication continued to influence treatment integrity implementation, $\chi^2(4, N = 144) = 4{,}153.87$, $p < .001$, $r^2 = .11$. Treatment integrity was addressed to a greater extent when skill-building treatments were evaluated, as compared with non-skill-building interventions, $t(192) = 2.14$, $p < .05$. Treatments in *JCCP*, *JAACAP*, and *AGP* were evaluated with greater attention to treatment integrity, as compared with treatments in *JCP, AJP,* and *BJP,* $t(140) = 2.64$, $p < .01$. The author's degree, $t(140) = 1.74$, *ns,* and country of degree, $t(140) < 1.00$, *ns,* were no longer significant predictors.

Discussion

Study 2 examined factors associated with the implementation of treatment integrity procedures. The results of the study indicated that skill-building approaches, relative to process-oriented and supportive–educational therapies, were evaluated with greater attention to treatment integrity. The specificity and concreteness of an intervention may at least partially explain this discrepancy. Skill-building approaches (e.g., cognitive–behavioral therapy) utilize specific techniques and strategies in alleviating psychological disturbances, whereas non-skill-building treatments (psychodynamic, existential therapies) place value on spontaneous responding in the moment and view a detailed treatment plan as detrimental (Bohart et al., 1998). Thus, skill-building approaches may be viewed as being less procedurally complex and as allowing more uniformity in behavior between therapists. Such uniformity permits the operational definition of an intervention, but improvisation, spontaneity, and creativity are more difficult to manualize.

The results also indicated that the journal of publication was related to treatment integrity implementation. Inquiry into the guidelines for authors in relation to reporting RTCs revealed no differences between journals. All of the evaluated journals required use of the Consolidated Standards of Reporting Trials (CONSORT; Moher, Schulz, & Altman, 2001). CONSORT presents a checklist of procedures to be performed during empirical tests and to be included in reports of the results. The checklist contains several items for integrity procedures, such as provision of detailed information on how the tested intervention was actually administered and how to report all departures from the protocol, and includes unplanned changes to interventions. However, these integrity procedures are loosely defined and leave much room for interpretation of the methods of implementation. Thus, although requirements are exactly the same, the outcome may depend on the authors' judgment of what should be done.

Training in research procedures may influence the authors' judgment. The results indicated that corresponding authors' educational background predicted the degree to which treatment integrity was addressed. That is, authors with research degrees

[5] We computed the r^2 value using the method articulated by Snijders and Bosker (1999). Currently, there is no clear consensus in the field on the ideal conceptualization and the computation of effect-size indices in a multilevel analytical context (Roberts & Monaco, 2006). Although several researchers have recommended computation of the percent-reduction in either the level-one or the level-two variances with the inclusion of the level-one or the level-two predictor variables, Snijders and Bosker have noted that it is not uncommon for such indices to be negative in value. Thus, Snijders and Bosker recommended an adjusted computational approach that is less likely to result in negative values of the proportion of explained variability. All r^2 values presented in the current text were computed with this method.

implemented more integrity procedures than did authors with nonresearch degrees. This effect may have influenced the relation between the journal of publication and treatment integrity implementation. Journals that published fewer studies with adequate implementation of integrity procedures included more articles by authors with nonresearch degrees (i.e., $r = .24$, $p < .001$ for *JCP* and $r = .27$, $p < .001$ for *BJP*). *JCCP*, on the other hand, contained more studies that adequately addressed integrity and published more articles by authors with research degrees ($r = .51$, $p < .001$). However, the results did not provide a consistent pattern, as the quality of publications in some journals (i.e., *AJP* and *AGP*) did not relate to authors' degree, and *JAACAP* actually published more articles by authors with nonresearch degrees ($r = .29$, $p < .001$). Confounding variables not controlled in this investigation, such as settings, may have affected this relationship.

Although the obtained effects of the predictor variables on treatment integrity were statistically significant, the effect size for treatment approach was small, and the effect size for journal was small to moderate. Overall, these two predictors accounted for approximately 11% of the variability in the implementation of treatment integrity. Thus, future research should investigate other potential predictors, such as treatment, therapist and client characteristics, and various barriers to implementation of treatment integrity procedures. We also should bear in mind that effect size is, in part, a function of the observed variability in the predictor and the criterion variables in a study. In the case of treatment integrity, adequate implementation of integrity procedures occurred infrequently, regardless of the value of various predictors, and this restricted variability in the criterion variable may have suppressed the magnitude of the effects. Thus, it will be beneficial to reevaluate the questions investigated in the current work, once adequate attention to treatment integrity is more widespread.

Several variables we expected would predict treatment integrity did not demonstrate a significant relationship. We predicted that in order to establish treatment differentiation, integrity procedures would be implemented to a higher degree when two or more active interventions were compared, relative to studies that evaluated only one active treatment. This expectation was not confirmed. Inadequate attention to integrity even when several interventions are evaluated may contribute to the Dodo bird effect, which refers to the claim that psychological treatments generally produce similar results (e.g., Wampold et al., 1997). Evidence exists to support and to challenge the equality notion (e.g., Crits-Christoph, 1997; Wampold et al., 1997). Whatever the final verdict in this debate, when comparisons of multiple therapies fail to produce differential effects of treatment, valid inferences cannot be drawn in regard to the outcomes unless treatment differentiation was established. Further, the results did not support our expectation that full reports, as compared with brief reports, would address integrity to a higher extent. However, the small number of brief reports included in the current study may have reduced statistical power.

General Discussion

Treatment integrity has important implications for the validity of the inferences drawn about the obtained effect. Although the methodological necessity of treatment integrity has long been recognized, few studies adequately implement treatment integrity procedures. Thus, guidelines for empirical testing of psychological treatments require reevaluation. Current concerns with the general approach to addressing treatment integrity can be illustrated by the criteria for demonstration of efficacy adopted by the American Psychological Association (APA) Division 12 Task Force on the Promotion and Dissemination of Psychological Procedures. The Task Force identified 16 distinct treatments as empirically supported and 56 interventions as probably efficacious (Chambless et al., 1998). These treatments shared four characteristics: skill-building techniques, specific focus, brief treatment contact, and continuous assessment of the dependent variable (O'Donohue, Buchanan, & Fisher, 2000). Although the goal in empirical testing is demonstration that unique treatment ingredients are responsible for the effect, almost no assessment of the delivery of these ingredients was included in the APA criteria. At best, utilization of a treatment manual and training of therapists were mentioned. The adequate implementation of integrity procedures was not enforced, which resulted in an identification of therapies as empirically supported primarily on the basis of whether changes on the dependent measures were observed.

Treatment integrity is not esoteric but rather is fundamental to empirical testing. Yet, enforcement of the implementation of treatment integrity procedures should be approached cautiously, as many questions pertaining to integrity require further elaboration. For example, our coding system suggests that several aspects of integrity should be evaluated in a study (e.g., therapist treatment adherence and competence). Nonetheless, it remains to be determined whether the obtained outcome or conclusions about treatment vary as a function of whether all or only a subset of the aspects of integrity is assessed. As an analogy, random assignment is invariably preferred to nonrandom assignment, and we take this as a given. However, randomization does not invariably lead to different conclusions (Shadish & Ragsdale, 1996). Although conceptually important, evaluation of all of the treatment integrity aspects may fail to enhance the incremental validity of our inferences when examined empirically.

Further, although optimal reporting of integrity would include all of its elements, researchers may choose a gradual, step-by-step approach to the evaluation of treatment integrity. For example, data on the validity and reliability of the integrity measures might be published separately from the primary treatment-outcome report. This possibility should be taken into account when one identifies treatments as empirically supported. The criteria for demonstrating efficacy should provide for the gradual evaluation and publication of data on the independent variable.

Implementation of treatment integrity procedures is costly and resource intensive, which almost certainly has deterred researchers from adequately addressing integrity. It is imperative to conduct the necessary cost–benefit analysis for determination of which integrity procedures must be implemented to ensure the validity of our conclusions. It may be the case that placing a higher premium on the validity of our inferences levies a price by permitting fewer studies. Integrity assessment is a matter of degree, and we need to know much more about the point at which further assessment is no longer beneficial.

Although the exact requirements for implementation of integrity procedures necessitate further elaboration, awareness of the issues should be raised. Awareness may be increased by organizing symposia on treatment integrity, devoting special sections of journals to theoretical and research papers on integrity implementation,

and soliciting conference presentations on the relevant issues. Removal of integrity data before publishing a study should be reexamined. Some authors or journal editors may choose to curtail information on treatment integrity in order to bring articles within page limits set forth by the publisher. However, sacrificing integrity data suggests to readers and researchers that treatment integrity is not important, indicates that controlling only one class of experimental variables (i.e., dependent measures) is sufficient, and renders questionable the validity of the inferences drawn from published research.

Further, journals may provide more precise specification of their requirements for reports of results of RCTs. CONSORT criteria for addressing treatment integrity are vague and need to be supplemented with concrete strategies. Davidson et al. (2003) suggest adding training and supervision of therapists, as well as assessment of treatment delivery, to CONSORT. Further, Borrelli et al. (2005) recommend including a fidelity framework for treatment design, training of therapists, delivery and receipt of treatment, and enactment of treatment skills. These suggestions point to a need for a more thorough attention to and report of integrity procedures. Yet, these recommendations should be considered cautiously, as inflexible criteria for the implementation of treatment integrity procedures may preclude gradual evaluation and publication of integrity data. Although conventional criteria for adequate attention to integrity are necessary, recommendations on specific procedures should take into account the need for a step-by-step approach for evaluation of treatment integrity.

This study examined RCTs published in influential psychiatric and psychological journals, which limits the generality of the findings. The external validity of the results was minimized in order to demonstrate that attention to treatment integrity is minimal even in articles published in gold-standard journals. Sample restriction may also have limited the obtained effects to experimental designs. However, if attention to treatment integrity is grossly inadequate in studies that represent the state of the art in research methods, the degree of implementation of treatment integrity procedures presumably is even lower in other types of treatment outcome research (e.g., those using more naturalistic designs). The inclusion of such studies would have severely constrained the range in integrity scores, such that moderate-to-high scores for integrity procedures would have been virtually absent. Such restriction in range would not have allowed adequate testing of potential predictors of variation in integrity implementation.

We selected the journals evaluated in this study on the basis of their impact factors to ensure that we included some of the most visible journals that are highly sought by researchers. However, an impact factor does not guarantee the highest quality studies or studies with a particular feature (e.g., integrity assessment). In fact, journals with lower impact scores might still house better studies. Consequently, a limitation of the paper is that it does not include a larger set of journals or journals with a lower impact factor and lower frequency of publication of the treatment outcome studies. Yet, the limited number of treatment outcome studies per journal (< 100) may have produced a restricted sample, which may not have allowed for comparisons between journals (e.g., *Behavior Therapy*, No. 94, by number of occurrences of treatment outcome studies; 38 studies within the specified years; listed between 15th and 35th by impact factor). Strict selection criteria resulted in an average of only 13.50% of treatment outcome studies selected from the identified articles. For example, although we found 479 treatment outcome articles within the specified years in *JCP*, only 6 studies (1.25%) satisfied our criteria for article selection.

The journals that were selected for this study are well-known outlets and serve as models for many investigators who conduct psychotherapy research. Our aim was to evaluate implementation of treatment integrity procedures in the gold standard of treatment outcome research, and our criteria for article selection reflect this goal. Yet, it may be informative to evaluate how treatment integrity is addressed in articles published in less influential journals, as compared with articles published in journals that have a higher impact factor. Such examination may provide us with more insights into the variables that affect treatment integrity implementation.

Further, the hierarchical relationship between integrity domains may have biased the results. That is, studies that fail to adequately establish integrity potentially run a risk of failure to adequately assess and evaluate integrity. Thus, authors may have been penalized several times for the same problem. On the other hand, low scores on implementation of procedures for one domain do not necessitate low scores for other domains. Provision of a general manual and failure to train and supervise therapists should not preclude the videotaping of every session and the utilization of valid and reliable integrity measures. Similarly, utilization of unvalidated measures should not affect adequate training of raters and assessment of interrater reliability. Indeed, data obtained in this study demonstrated that report of treatment integrity was more adequate than were assessment and evaluation of integrity.

Authors of the articles included in the present analyses were not contacted for information in regard to the implemented treatment integrity procedures. Contacting authors might have supplied additional data, which might have influenced our conclusions. That is, researchers might have reported that they implemented more integrity procedures than were detailed in the articles. Yet, the reliability of such reports would be impossible to evaluate, as self-reports are subject to distortions and poor recollection. Further, contacting authors for this information might have resulted in high levels of missing data that would have biased the results.

Our decision to rely on published data on treatment integrity instead of to contact authors restricts our knowledge about the specific procedures utilized for measurement of treatment differentiation. Ideally, therapist treatment adherence measures include prescribed as well as proscribed tasks (i.e., procedures to avoid), which is sufficient to establish that treatments are distinct. Yet, without consultation of each treatment adherence measure, our confidence that both types of procedures were monitored is limited. Thus, it is hard for us to estimate how treatment differentiation was addressed in our sample. The establishment of treatment differentiation is critical, especially given the Dodo bird effect. The diluted interventions may attenuate the magnitude of the obtained effect or may fail to show differential effects on dependent measures, because treatment components are no longer distinct. Manipulation checks on the proscribed procedures should be included in adherence measures in order to ensure treatment purity, and description of utilized procedures should be reported.

Future research may examine how treatment differentiation is addressed in the psychotherapy outcome literature. This task may entail careful perusal of therapist treatment adherence measures for examination of the extent to which researchers check for pro-

scribed procedures along with prescribed tasks. Such examination may increase awareness of the importance of treatment integrity and may provide guidelines for ensuring adequacy of the implemented procedures.

Addressing treatment integrity is expensive and laborious. Future research may evaluate ways to attain a satisfactory balance of the costs and benefits of attending to integrity. For example, how can one best achieve accurate representation of integrity data? What are the optimal number and length of observations? Validation of integrity measures represents a particular challenge, as such measures may encompass two separate constructs—therapist treatment adherence and therapist competence. Additionally, treatments may differ in their operational definitions, components, and requirements for competent implementation. Does this mean that integrity measures may have to be developed and validated specifically for each treatment, or can ways be devised to create more general measures of integrity? Further empirical examination of the relation between the different aspects of treatment integrity might provide a much-needed insight into the question of the incremental utility of evaluating various aspects of integrity.

The empirical evaluation of process-oriented psychotherapies warrants greater consideration. With the increased use of empirically supported interventions in clinical practice and training, these potentially efficacious treatments may become obsolete just because they resist empirical testing with the current methods. Indeed, the predominant majority of validated treatments—85%—are skill-building interventions (O'Donohue et al., 2000).

Future research might examine additional predictors of integrity implementation. Such predictors might include treatment, therapist and client characteristics, various barriers to implementation of treatment integrity procedures (e.g., lack of editorial requirement), and characteristics of the journal of publication (e.g., specific vs. general scope, high vs. low impact factor, psychiatric vs. psychological). Future research might also evaluate whether studies that more adequately address integrity procedures and achieve higher integrity levels are more likely to find significant differences between therapies. A demonstration that attending to treatment integrity advances science, rather than merely meets a reporting requirement, may serve as a powerful incentive for researchers to adhere to treatment integrity regulations.

References

Addis, M. E., Wade, W. A., & Hatgis, C. (1999). Barriers to dissemination of evidence-based practices: Addressing practitioners' concern about manual-based psychotherapies. *Clinical Psychology: Science and Practice, 6,* 430–441.

Barber, J. P., Gallop, R., Crits-Christoph, P., Frank, A., Thase, M. E., Weiss, R. D., & Connolly Gibbons, M. B. (2006). The role of therapist adherence, therapist competence, and alliance in predicting outcome of individual drug counseling: Results from the National Institute Drug Abuse Collaborative Cocaine Treatment Study. *Psychotherapy Research, 16,* 229–240.

Bergan, J., & Kratochwill, T. (1990). *Behavioral consultation and therapy.* New York: Plenum Press.

Binder, J. (1993). Observations on the training of therapists in time-limited dynamic psychotherapy. *Psychotherapy, 30,* 592–598.

Bohart, A. C., O'Hara, M., & Leitner, L. M. (1998). Empirically violated treatments: Disenfranchisement of humanistic and other psychotherapies. *Psychotherapy Research, 8,* 141–157.

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Breger, R., DeFrancesco, C., Levesque, et al. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology, 73,* 852–860.

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

Carroll, K. M., & Nuro, K. F. (2002). One size cannot fit all: A stage model for psychotherapy manual development. *Clinical Psychology: Science and Practice, 9,* 396–406.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., et al. (1998). Update on empirically validated therapies, II. *Clinical Psychologist, 51,* 3–16.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 1–18.

Codding, R. S., Feinberg, A. B., Dunn, E. K., & Pace, G. M. (2005). Effects of immediate performance feedback on implementation of behavior support plans. *Journal of Applied Behavior Analysis, 38,* 205–219.

Crits-Christoph, P. (1997). Limitations of the Dodo bird verdict and the role of clinical trials in psychotherapy research: Comment on Wampold et al. (1997). *Psychological Bulletin, 122,* 216–220.

Davidson, K. W., Goldstein, M., Kaplan, R. M., Kaufmann, P. G., Knatterud, G. L., et al. (2003). Evidence-based behavioral medicine: What is it and how do we achieve it*? Annals of Behavior Medicine, 26,* 161–171.

Drozd, J. F., & Goldfried, M. R. (1996). A critical evaluation on the state-of-the-art in psychotherapy outcome research. *Psychotherapy, 33,* 171–180.

Gresham, F. M. (1989). Assessment or treatment integrity in school consultation and prereferral interventions. *School Psychology Review, 18,* 37–50.

Gresham, F. M. (1997). Treatment integrity in single-subject research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 93–117). Mahwah, NJ: Erlbaum.

Gresham, F. M., Donald, L., MacMillan, D. L., Beebe-Frankenberger, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice, 15,* 198–205.

Gresham, F. M., Gansle, K. A., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis, 26,* 257–263.

Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S., & Rosenblum, S. (1993). Treatment integrity in school-based intervention studies: 1980–1990. *School Psychology Review, 22,* 254–272.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques.* Mahwah, NJ: Erlbaum.

Henry, W. P. (1998). Science, politics, and the politics of science: The use and misuse of empirically validated treatments. *Psychotherapy Research, 8,* 126–140.

Johnston, J., & Pennypacker, H. S. (1980). *Strategies and tactics of human behavioral research.* Hillsdale, NJ: Erlbaum.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

Kendall, P. C., & Chu, B. C. (2000). Retrospective self-reports of therapist flexibility in a manual-based treatment for youths with anxiety disorders. *Journal of Clinical Child Psychology, 29,* 209–220.

Kratochwill, T. R., Elliott, S. N., & Busse, R. T. (1995). Behavioral consultation: A five-year evaluation of consultant and client outcomes. *School Psychology Quarterly, 10,* 87–117.

McGlinchey, J. B., & Dobson, K. S. (2003). Treatment integrity concerns in cognitive therapy for depression, *Journal of Cognitive Psychotherapy: An International Quarterly, 17,* 299–318.

Moher, D., Schulz, K. F., & Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of

parallel-group randomized trials. *Journal of the American Medical Association, 285,* 1987–1991.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247–266.

Nezu, A. M., & Nezu, C. M. (2005). Comments on "Evidence-based behavior medicine: What it is and how do we achieve it?": The interventionist does not always equal the intervention—the role of therapist competence. *Annals of Behavior Medicine, 29,* 80.

O'Donohue, W., Buchanan, J. A., & Fisher, J. E. (2000). Characteristics of empirically supported treatments. *Journal of Psychotherapy Practice and Research, 9,* 69–74.

Paivio, S. C., Holowaty, K. A. M., & Hall, I. (2004). The influence of therapist adherence and competence on client reprocessing of child abuse memories. *Psychotherapy: Theory, Research, Practice, Training, 41,* 56–68.

Patterson, G. R., & Chamberlain, P. (1994). A functional analysis of resistance during parent training therapy. *Clinical Psychology: Science and Practice, 1,* 53–70.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12,* 365–383.

Peterson, L., Homer, A., & Wonderlich, S. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15,* 477–492.

Roberts, J. K., & Monaco, J. P. (2006, April). *Effect size measures for the two-level linear multilevel model.* Paper presented at the annual meeting of the American Education Research Association.

Rogers Wiese, M. R. (1992). A critical review of parent training research. *Psychology in the School, 29,* 229–236.

Schlosser, R. W. (2002). On the importance of being earnest about treatment integrity. *Augmentative and Alternative Communication, 18,* 36–44.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments. Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Shaw, B. F., Elkin, I. E., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobsson, K. S., et al. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67,* 837–846.

Snijders, T., & Bosker, N. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

Sterling-Turner, H. E., Watson, T. S., & Moore, J. W. (2002). The effects of direct training and treatment integrity on treatment outcomes in school consultation. *School Psychology Quarterly, 17,* 47–77.

Sterling-Turner, H. E., Watson, T. S., Wildmon, M., Watkins, C., & Little, E. (2001). Investigating relationship between training type and treatment integrity. *School Psychology Quarterly, 16,* 56–67.

Tabachnick, B. G., & Fidell, S. L. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Thomson ISI. (2002). *Journal Citation Reports.* Retrieved January 12, 2005, from www.isinet.com/products/evaltools/jcr/

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). San Diego, CA: Academic Press.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61,* 620–630.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "all must have prizes." *Psychological Bulletin, 122,* 203–215.

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130,* 631–663.

Wilson, G. T. (1998). Manual-based treatment and clinical practice. *Clinical Psychology: Science and Practice, 5,* 363–375.

Received June 21, 2006
Revision received May 31, 2007
Accepted June 4, 2007 ■