Journal of Consulting and Clinical Psychology
2005, Vol. 73, No. 5, 852–860

0022-006X/05/$12.00 DOI: 10.1037/0022-006X.73.5.852

A New Tool to Assess Treatment Fidelity and Evaluation of Treatment Fidelity Across 10 Years of Health Behavior Research

Belinda Borrelli and Deborah Sepinwall
Brown Medical School and Miriam Hospital

Denise Ernst
University of New Mexico

Albert J. Bellg
Appleton Heart Institute

Susan Czajkowski
National Institutes of Health

Rosemary Breger and Carol DeFrancesco
Oregon Health Sciences University

Chantal Levesque
Southwest Missouri State University

Daryl L. Sharp
University of Rochester

Gbenga Ogedegbe
Columbia University

Barbara Resnick and Denise Orwig
University of Maryland

A. Bellg, B. Borrelli, et al. (2004) previously developed a framework that consisted of strategies to enhance treatment fidelity of health behavior interventions. The present study used this framework to (a) develop a measure of treatment fidelity and (b) use the measure to evaluate treatment fidelity in articles published in 5 journals over 10 years. Three hundred forty-two articles met inclusion criteria; 22% reported strategies to maintain provider skills, 27% reported checking adherence to protocol, 35% reported using a treatment manual, 54% reported using none of these strategies, and 12% reported using all 3 strategies. The mean proportion adherence to treatment fidelity strategies was .55; 15.5% of articles achieved greater than or equal to .80. This tool may be useful for researchers, grant reviewers, and editors planning and evaluating trials.

Keywords: treatment fidelity, clinical trials, internal validity, health behavior change.

Belinda Borrelli and Deborah Sepinwall, Centers for Behavioral and Preventive Medicine, Brown Medical School, and Miriam Hospital, Providence, Rhode Island; Denise Ernst, Department of Psychology, Center on Alcoholism, Substance Abuse, and Addiction, University of New Mexico; Albert J. Bellg, Appleton Heart Institute; Susan Czajkowski, Behavioral Medicine Scientific Research, National Heart, Lung, and Blood Institute, National Institutes of Health; Rosemary Breger and Carol DeFrancesco, Division of Health Promotion and Sports Medicine, Oregon Health Sciences University; Chantal Levesque, Department of Psychology, Southwest Missouri State University; Daryl L. Sharp, School of Nursing, University of Rochester; Gbenga Ogedegbe, Behavioral Cardiovascular Health and Hypertension Program, Department of Medicine, Columbia University; Barbara Resnick, School of Nursing, University of Maryland; Denise Orwig, Department of Epidemiology and Preventive Medicine, School of Medicine, University of Maryland.

Albert J. Bellg is now at the Department of Psychology, Lawrence University.

This research was supported in part by an administrative supplement from the National Heart, Lung, and Blood Institute, awarded to Belinda Borrelli (Grant R01 HL62165).

Correspondence concerning this article should be addressed to Belinda Borrelli, Center for Behavioral and Preventive Medicine, Brown Medical School, Coro Building—West, 1 Hoppin Street, Suite 500, Providence, RI 02903. E-mail: belinda_borrelli@brown.edu

Treatment fidelity refers to the methodological strategies used to monitor and enhance the reliability and validity of behavioral interventions. The overall goal of enhancing treatment fidelity is to increase scientific confidence that changes in the dependent variable are attributable to the independent variable. Careful consideration of treatment fidelity helps to explain study findings, revise interventions for future testing, and increase statistical power and effect size by reducing random and unintended variability (Moncher & Prinz, 1991). Enhancing treatment fidelity has the effect of not only increasing internal validity but also increasing external validity, as a high degree of treatment fidelity is needed both for study replication and for generalization of treatments to applied settings. For treatments to be adopted by clinicians and/or integrated into existing infrastructures, information about method, fidelity, and effectiveness is needed. The cost of inadequate fidelity can be rejection of powerful treatment programs or acceptance of ineffective programs (Henggeler, Melton, Brondino, Scherer, & Hanley, 1997; Moncher & Prinz, 1991).

The concept of treatment fidelity and the strategies involved in maintaining treatment fidelity have broadened significantly in the last 20 years. Starting with the original idea of *treatment integrity* (whether the treatment was delivered as intended), various researchers added *treatment differentiation* (whether the treatments

differ in the intended manner; Kazdin, 1986; Moncher & Prinz, 1991; Yeaton & Sechrest, 1981), *treatment receipt* (whether the client comprehends and uses the treatment skills during the session), and *treatment enactment* (whether the client actually applies skills learned in treatment to his or her daily life, between sessions; Burgio et al., 2001; Lichstein, Riedel, & Grieve, 1994). This expansion of the concept of treatment fidelity was necessary as effectiveness trials, hybrid efficacy–effectiveness trials, and patient–treatment matching research became more prevalent (Carroll et al., 1998).

More recently, a comprehensive five-part treatment fidelity framework has been developed that integrates previous conceptualizations of treatment fidelity, adds some novel components, and is tailored to be relevant for health behavior change research and clinical practice (Bellg, Borrelli, et al., 2004; Borrelli et al., 2002). This framework adds to the previous conceptualizations of treatment fidelity by (a) including factors to consider both when designing a study and when training providers; (b) expanding on the above-mentioned areas of delivery, receipt, and enactment; and (c) increasing their relevancy to health behavior change trials. The treatment fidelity guidelines described in this article were developed by the Treatment Fidelity Workgroup, whose members were part of the National Institutes of Health (NIH) Behavioral Change Consortium (BCC). The BCC provided an infrastructure to support collaboration among 15 NIH-funded health behavior change projects. Preservation of treatment fidelity in these projects was critical, because all of the projects involved theory testing and the majority were conducted in real-world settings. The Treatment Fidelity Workgroup was established to advance the definition and measurement of treatment fidelity within the BCC and in health behavior change studies in general.

The workgroup developed a set of guidelines and recommendations for best practices that cover five categories: Design, Training, Delivery, Receipt, and Enactment. The *Design* category consists of factors that should be considered when designing a trial but also includes factors that should be reported in order to evaluate and replicate the trial. Some examples of such factors are information about content and dose for both the treatment and comparison conditions (length of each contact, number of contacts, duration of contact over time), information on the type of provider background needed to successfully implement the intervention (credentials, experience), and articulation of a theoretical framework or clinical guidelines on which the intervention is based.

The *Training* category of the treatment fidelity framework describes a number of issues that are important to consider for interventions that use human providers. Prior to training providers, investigators should think about the specific competencies required for the successful delivery of the intervention and develop the training accordingly. At this early stage, it is also important to hire those who are not only capable of delivering the intervention but also buy into the theoretical foundation. For example, one would not want to hire counselors who believe in "abstinence only" for effective alcoholism treatment if the protocol calls for cognitive–behavioral relapse prevention. Information about how the providers were trained, whether training was standardized across providers, measurement of provider skill acquisition, and how provider skills were maintained over time is also an important aspect of training that should be implemented and reported.

The third category of the framework, *Treatment Delivery*, focuses on processes that monitor and improve the delivery of the intervention so it can be established that the intervention was delivered as intended. Methods need to be included that increase the likelihood that both the content and dose of the intervention are being delivered as originally conceptualized (e.g., through use of a treatment manual and recording the amount of contact time). In addition, there should be a mechanism by which investigators are able to assess whether the provider adhered to the intervention plan (e.g., through audiotaped sessions). It is also important to measure nonspecific effects, to ensure that therapeutic alliance is similar across conditions.

The category of *Treatment Receipt* involves ensuring that participants understand the information provided in the intervention. This is especially important when participants are cognitively compromised or have low levels of literacy, education, or proficiency in English. Providers need to assess that participants are able to use the cognitive skills taught in the intervention (e.g., relapse prevention skills, problem solving) as well as the behavioral skills (e.g., how to use nicotine gum, relaxation techniques, food diaries) learned in session. *Treatment Enactment* consists of processes to monitor and improve the ability of patients to perform treatment-related cognitive strategies and behavioral skills in their daily lives (e.g., fills a pill organizer, uses a cognitive strategy to deal with craving for cigarettes, uses nicotine gum).

These treatment fidelity categories are mutually exclusive. Inattention to one category could compromise the internal validity of the study despite adherence in the other categories. For example, without assessing provider skill acquisition and maintenance, it cannot be determined whether nonsignificant results are due to an ineffective intervention or to lack of attention to these training issues. Heterogeneity in provider skills can result in Provider $\times$ Treatment interactions. Alternatively, providers may be well trained, but the intervention may not be delivered as it was intended. Without monitoring treatment delivery, significant results could be due to unknown active components added to the intervention, inactive ingredients omitted from the intervention (that had the function of diluting the intervention), or to an effective intervention (Moncher & Prinz, 1991). Alternatively, nonsignificant results may be due to the omission of active intervention ingredients, the addition of intervention contaminants, or to an ineffective intervention. It is difficult to isolate any one of these potentially causal factors without monitoring the degree to which the intervention was delivered as intended. Similarly, although there may be high levels of protocol adherence, without assessing client "receipt," researchers may falsely conclude that the intervention was not effective, when really the client did not understand how to perform the cognitive or behavioral skills learned in session. Finally, even if treatment receipt is established (e.g., the client understands how to do a relaxation procedure or how to use nicotine gum), the client may never attempt to try the intervention in between sessions, leading the provider to falsely conclude that the treatment is not working in the light of poor treatment response (e.g., high anxiety levels; high degree of nicotine craving). Others have also commented on the threat to validity caused by inattention to treatment receipt (Kazdin, 1986; Smith & Sechrest, 1991). Specific strategies and recommendations for promoting treatment fidelity in each of these areas are described in Bellg, Borrelli, et al. (2004).

The purpose of this study was twofold: (a) to develop an assessment tool based on our framework to help researchers evaluate the degree of treatment fidelity in their own and other's work and (b) to use the assessment tool to evaluate how all five components of treatment fidelity in our framework (Design, Training, Delivery, Receipt, and Enactment) have been addressed over the past 10 years in five key journals that publish health behavior change research.

Two prior studies have evaluated the occurrence of adherence to treatment fidelity in the extant literature. Moncher and Prinz (1991) assessed treatment fidelity in 359 studies published between 1980 and 1988 in journals from clinical psychology, behavior therapy, psychiatry, and marital and family therapy. Lichstein et al. (1994) assessed delivery, receipt, and enactment among articles published in two journals (*Journal of Consulting and Clinical Psychology* and *Behavioral Therapy*) for 1 full year (1990). The present study expands on these prior studies by (a) assessing the presence of treatment fidelity among articles published in five journals, (b) including studies published over a 10-year time span, (c) focusing only on health behavior change studies, and (d) expanding the assessment of treatment fidelity to include a comprehensive constellation of treatment fidelity indicators as described in our framework. We believe that a replication and extension of prior work is needed to identify areas for improvement in the health behavior change literature.

Method

We evaluated treatment fidelity practices as reported in the health behavior change literature between 1990 and 2000. Studies included in this review were identified through a hand search of five journals that are major publication outlets for health behavior change research: *Annals of Behavioral Medicine, Health Psychology, American Journal of Health Promotion, American Journal of Public Health,* and *Journal of Consulting and Clinical Psychology*.

Article inclusion and exclusion criteria were similar to those used by Moncher and Prinz (1991) in order to facilitate comparison with this earlier study. We included articles that were psychosocial interventions designed to treat a specific problem (e.g., cognitively based treatments, social skills training, behavioral interventions, and the like). Studies were excluded if they were (a) general interventions, such as attendance at meetings; (b) interventions directed at changing a health-related condition (e.g., high cholesterol) rather than a health behavior (e.g., diet); (c) analog studies; (d) evaluations of different methods for publicly disseminating information; (e) evaluations of training procedures that were not directed toward remediation of a specific health problem; or (f) focused on policy. The articles had to contain an experimental manipulation of treatment and report of data. Our eligibility criteria diverged from Moncher and Prinz (1991) in that we included single-group designs as well as quasi-experimental studies because we believed that treatment fidelity should be reported in these articles as well.

A total of 371 articles met inclusion criteria upon initial screening by Deborah Sepinwall. As members of the workgroup reviewed studies that were assigned to them for coding, 29 additional articles were deemed inappropriate on the basis of our exclusion criteria, resulting in a final sample of 342 articles. The health behaviors targeted for change by these studies included smoking, weight loss, nutrition, physical activity, alcohol and drug use, and safe sex. Some less frequent health behavior change studies represented in our data were seat belt adherence, sun safety, and using dental fluoride.

Characteristics of the Treatment Fidelity Coders

A central coordinating site was responsible for selecting and distributing articles to six pairs of coders (including themselves) across the country. Eight of the coders held doctoral degrees, three had master's degrees, and one was a physician. Eleven coders were university-affiliated, and one was an employee of the NIH. Their areas of specialization included psychology, medicine, public health, epidemiology, and nursing. All coders were members of the NIH BCC's Treatment Fidelity Workgroup and were involved in their own BCC studies.

Development of the Treatment Fidelity Checklist and Reliability of Coding Procedure

The checklist contains the list of criteria (25 items) by which articles were evaluated (see Table 1). The 25 items are divided into the five treatment fidelity categories (Design, Training, Delivery, Receipt, and Enactment). The original checklist was developed by Belinda Borrelli and Albert J. Bellg from several sources (e.g., survey of their own and other BCC sites for treatment fidelity practices, survey of the extant literature). The checklist was pilot tested in the BCC and further refined through pilot-test coding, during which all authors independently rated the same 10 articles. The percent agreement was computed between all coders to establish intercoder reliability, as recommended by Lombard, Snyder-Duch, and Bracken (2002). We achieved an 84% agreement between coders.

Coders indicated the presence or absence of a number of characteristics important for the preservation of treatment fidelity within each of the five categories. Each of the treatment fidelity items was followed by a series of examples to help the coder determine whether the article met the criteria for that particular item (available upon request). Treatment fidelity information was judged to be either "present" (i.e., the article mentioned use of a particular treatment fidelity strategy), "absent but should be present" (treatment fidelity information was inappropriately omitted, preventing the coder from being able to accurately assess the scientific validity of the article) or "not applicable" (the particular treatment fidelity strategy was not applicable to the article in question; e.g., articles using computer-tailored reports as an intervention were not penalized for not adhering to the strategies in the training category, because there were no human interventionists to be trained).

Process of Coding the Articles

The coding of articles began in July 2002 and ended in May 2003. A coding manual, which included definitions of the treatment fidelity categories and examples, was developed and used by the coders. Articles were distributed to pairs of coders. Articles were first coded individually, and then each coding pair met by telephone or in person to discuss their ratings on the same articles and resolve any discrepancies. During this meeting, the pair generated a uniform coding sheet to be submitted to the central coordinating site. Monthly conference calls were held with the full group to discuss coding issues and to clarify discrepancies.

Of the coded articles, 21% referred readers to previously published articles for more details about the treatment fidelity of their study. Our workgroup decided to pursue these referenced articles (though they were not in our targeted list of journals), code them, and give credit to the targeted article in the spirit of obtaining a fair assessment of the degree of treatment fidelity reported by that particular study. In other words, if we found low levels of treatment fidelity, we wanted to rule out the possibility that treatment fidelity for a particular study was reported in another journal other than the five we had targeted, because it is common practice that authors refer readers to other articles for additional information. Only one checklist was generated per original article assigned (even though more than one article may have been consulted for coding). The checklist was constructed so that coders could indicate whether the treatment fidelity

Table 1
Percentage of Articles Reporting Use of Treatment Fidelity Strategies

Treatment fidelity strategies	Targeted journals	Targeted journals + referenced articles	*n*
Treatment design			
1. Provided information about treatment dose in the intervention condition			
Length of contact session(s)	63.0%	67.0%	329
Number of contacts	86.0%	90.0%	331
Content of treatment	94.0%	98.0%	341
Duration of contact over time	92.0%	95.0%	341
2. Provided information about treatment dose in the comparison condition			
Length of contact session(s)	64.0%	67.0%	240
Number of contacts	83.0%	86.0%	246
Content of treatment	90.0%	93.0%	266
Duration of contact over time	87.0%	90.0%	252
3. Mention of provider credentials	64.0%	69.0%	293
4. Mention of a theoretical model or clinical guidelines on which the intervention is based	71.0%	74.0%	338
Training providers			
1. Description of how providers were trained	25.0%	29%	291
2. Standardized provider training	25.0%	29%	291
3. Measured provider skill acquisition posttraining	16.0%	18%	292
4. Described how provider skills maintained over time	22.0%	27%	288
Delivery of treatment			
1. Included method to ensure that the content of the intervention was being delivered as specified (e.g., treatment manual, checklist, computer program)	46.0%	51.0%	332
2. Included method to ensure that the dose of the intervention was being delivered as specified (e.g., records number of contact minutes)	31.0%	34.0%	322
3. Included mechanism to assess if the provider actually adhered to the intervention plan (applies to human providers only?) (e.g., audiotape, observation, self-report of provider, exit interview with participant)	27.0%	30.0%	287
4. Assessed nonspecific treatment effects	6.0%	7.0%	283
5. Used treatment manual	35.0%	38.0%	301
Receipt of treatment			
1. Assessed subject comprehension of the intervention during the intervention period	40.0%	45.0%	332
2. Included a strategy to improve subject comprehension of the intervention above and beyond what is included in the intervention	52.0%	57.0%	331
3. Assessed subject's ability to perform the intervention skills during the intervention period	50.0%	54.0%	326
4. Included a strategy to improve subject performance of intervention skills during the intervention period	53.0%	58.0%	325
Enactment of treatment skills			
1. Assessed subject performance of the intervention skills assessed in settings in which the intervention might be applied	69.0%	73.0%	330
2. Assessed strategy to improve subject performance of the intervention skills in settings in which the intervention might be applied	46.0%	50.0%	327

information was obtained from the primary (i.e., originally retrieved article) or secondary (previously published article) source. This differentiation allowed us to analyze the data with and without the additionally referenced article.

Maintaining Intercoder Reliability Over Time

A third rater (Deborah Sepinwall) coded 20% ($n = 68$) of the studies coded by the pairs, an amount that is twice the current recommendation (Lombard et al., 2002). The percent agreement was calculated by the ratio of the number of discrepancies divided by the number of items for each article. Percent agreement ranged from 77% to 96% ($M = 87\%$). We acknowledge that percent agreement is a more liberal index of intercoder reliability, but other indices, such as kappa, may be overly conservative given the nature and purpose of our instrument (Lombard et al., 2002). Percent agreement may be used if the cutoff for acceptability is set high, such as .80 or greater (Lombard et al., 2002). Discrepancies between coders were addressed in the same manner as was done within the pairs (i.e., through discussion). Coders met monthly by telephone to discuss coding problems and questions.

Analytic Plan

The percentage of articles using a particular strategy was computed by the ratio of the number of articles that coders deemed as using that particular strategy to the total number of articles for which the strategy was considered appropriate. Therefore, if the particular strategy was not applicable to an article's study design (e.g., training providers would not be relevant for an intervention delivered by computer), that study was not

included in the denominator. All analyses on the individual strategies were conducted with chi-square statistics.

Next, as a way of aggregating the data, the mean proportion of adherence to treatment fidelity strategies was calculated for each individual article by summing the number of strategies coded as "present" and dividing by the number of strategies coded as appropriate for that study design. The mean proportion of adherence to treatment fidelity strategies was then summed across articles and means were computed for each category (e.g., Design, Training, Delivery, Receipt, Enactment) as well as across all categories (index of overall mean treatment fidelity). Between-group differences on continuous variables were examined with analysis of variance (ANOVA; PROC ANOVA, SAS system for Windows, Version 8.0). Changes over time in treatment fidelity strategies were conducted with repeated measures ANOVAs. Tukey's honestly significant difference test was used to examine univariate contrasts.

Results

The total number of articles included in our study was 342. The distribution of articles across journals was as follows: 129 articles (38%) in the *American Journal of Public Health,* 96 (28%) in the *Journal of Consulting and Clinical Psychology,* 50 (15%) in *Health Psychology,* 49 (14%) in the *American Journal of Health Promotion,* and 18 (5%) in the *Annals of Behavioral Medicine*.

Frequency of Reporting Treatment Fidelity Strategies

Table 1 displays the percentage of articles using each of the treatment fidelity strategies. The "Targeted journals" column displays the percentage of articles that included the treatment fidelity strategy when only articles published in the five targeted journals were considered. A total of 71 articles (21%) from the five targeted journals required that we retrieve articles from another journal because the referenced article included treatment fidelity strategies referred to but not expounded on in the targeted article. Therefore, the "Targeted journals + referenced articles" column contains the percentage of articles that included the treatment fidelity strategy when articles in our five targeted journals as well as these referenced articles were considered. Thus, this column is a less conservative estimate of the presence of a particular treatment fidelity strategy, in that it gives credit to articles that were in one of the targeted journals and referred the reader to an earlier article for a full explanation of treatment fidelity. Comparison between these two columns in Table 1 reveals that inclusion of the referenced articles improved treatment fidelity by 3% to 6% (mean improvement = 4.3%) over the percentage found when we included articles only from the targeted journals. Overall, the percent use of treatment fidelity strategies varied greatly by category. Items within the individual categories ranged from 63% to 94% in the Study Design category, 16%–25% in the Training category, 6.0%−46% in the Delivery category, 40%–53% in the Receipt category, and 46%–69% in the Enactment category.

Frequency of Reporting Treatment Fidelity Strategies and Changes Over Time: Comparison With Moncher and Prinz (1991)

We more closely examined three of the treatment fidelity strategies listed in Table 1 because they were also examined by Moncher and Prinz (1991). These strategies were supervision of treatment providers, checking adherence to protocol, and use of a treatment manual. Moncher and Prinz found that 31.5% of studies used a treatment manual, whereas we found that 35% (105/301) studies did so; they found that 21.4% of studies supervised treatment agents, whereas we found that 22% (63/288) did so. We did, however, find a higher percentage of studies checking adherence to protocol (27%; 77/287) than did Moncher and Prinz (18.1%). Moncher and Prinz found that 55.3% of their studies used none of these three strategies, whereas we found that 54% of health behavior change studies used none of the three strategies. Alternatively, we found that 12% of studies used all three of these strategies, compared with Moncher and Prinz (1991), who reported that only 5.8% of their studies used all three strategies. In our study, the use of these strategies did not significantly change over time (see Table 2). There were, however, nonsignificant trends for decreases in reporting from the early 1990s to the late 1990s ($p < .10$). Moncher and Prinz, on the other hand, found significant increases over time in the percentage of articles that reported supervision of treatment agents and checking adherence to the protocol, as well as a threefold increase in the percentage of studies that reported using all three treatment fidelity procedures. The use of treatment manuals did not significantly change over time in the Moncher and Prinz study.

Mean Proportion Adherence to Treatment Fidelity Strategies Grouped by Category

We examined whether the mean proportion adherence of the articles to treatment fidelity strategies was greater for some treatment fidelity categories versus others (see Table 3). Across all

Table 2
Frequency of Reporting Treatment Fidelity Strategies Over Time

	1990–1993		1994–1997		1998–2000	
Treatment fidelity item	%	*n*	%	*n*	%	*n*
Provider skill maintenance over time	25	69	19	118	23	101
Mechanism to assess provider adherence to protocol	32	71	25	117	25	99
Use of a treatment manual	33	78	39	117	31	106
Use of all 3 strategies	15	69	11	114	11	98
Use of none of the 3 strategies	55	69	50	114	58	98

Note. These strategies were extracted from Table 1 because they were similar to Moncher and Prinz (1991), therefore facilitating comparison.

Table 3
Proportion of Adherence to Treatment Fidelity Strategies in Articles in Targeted Journals

Category	Mean proportion	Median	*SD*	Number of articles
Design	.80	.86	.22	342
Training	.22	.00	.34	292
Delivery	.35	.33	.33	334
Receipt	.49	.50	.38	337
Enactment	.57	.50	.40	331
Mean overall adherence	.55	.56	.22	342

articles from the targeted journals, the average proportion of adherence to treatment fidelity strategies within the Design category was .80. The lowest mean proportion of adherence to strategies was found in the Training category, where, on average, only .22 of strategies were reported among applicable studies. The mean proportion of adherence to strategies in the Delivery, Receipt, and Enactment categories was .33, .49, and .57, respectively.

We calculated the overall proportion of adherence to the strategies by summing for each article the total number of individual items on the checklist coded as "present" divided by the total number of individual items that were considered applicable to the study. On average, when all of the articles from targeted journals were included ($N = 342$), the mean proportion of adherence to treatment fidelity strategies included on our coding sheet was .55; when the information from the referenced articles was counted in the proportion adherence, the proportion adherence increased significantly to .59, $t(342) = 5.55$, $p = .0001$. If we looked separately at the articles that did not refer the reader to another article for treatment fidelity information ($n = 271$), the overall mean proportion adherence to the strategies was .60. For the 71 articles that did refer the reader to another article for treatment fidelity, the overall mean proportion adherence to the strategies was .34 without including the referenced articles, and .53 when the additional article was included.

High Levels of Treatment Fidelity

We defined "high treatment fidelity" as those studies that had .80 or greater proportion adherence to our checklist across all strategies. A total of 15.5% (53 out of 342) of articles from the targeted journals had .80 or greater adherence. This was calculated by dividing the number of strategies used by the study by the number of strategies applicable to the study design. The percentage of articles that achieved .80 adherence or greater in each category was as follows: Design, 68% (231 of 342); Training, 10% (30 of 292); Delivery, 20% (68 of 334); Receipt, 23% (78 of 337); and Enactment, 42% (138 of 331). Twenty-two studies (6.5%) achieved .80 adherence across all strategies and all categories. Examples of such studies are Cameron et al. (1999); Maude-Griffin et al. (1998); Stephens, Roffman, and Curtin (2000); Fals-Stewart, Birchler, and O'Farrell (1996); and Rotheram-Borus, Reid, and Rosario (1994).

Discussion

The purpose of the present study was to (a) provide researchers with an assessment tool to evaluate treatment fidelity in their own and other's trials and (b) use the assessment tool to report on treatment fidelity practices across a 10-year time span in five major journals that publish health behavior change trials. Our treatment fidelity assessment tool was extensively piloted and demonstrated a high degree of intercoder reliability. We used this tool to determine (a) the percentage of articles that used each strategy and (b) the proportion of strategies reported by each article. We also assessed change in the use of strategies over time. To our knowledge, this is the first study that assessed the degree of treatment fidelity among health behavior change trials.

The percent use of treatment fidelity strategies varied greatly by category. Items within the individual categories ranged from 63% to 94% in the Study Design category, 16%–25% in the Training category, 6.0%−46% in the Delivery category, 40%–53% in the Receipt category, and 46%–69% in the Enactment category. In comparing our findings with those reported by Moncher and Prinz (1991) on the three treatment fidelity strategies in common, there were remarkably similar percentages despite the fact that they reviewed a different literature (general clinical psychology) over a different time period (1980–1988). We expected to find improvements in treatment fidelity over time, in line with previous findings from Moncher and Prinz, but there was only a nonsignificant trend indicating a decrease in the use of these strategies over time.

In our study, the strategies in the training category had the lowest percentages reported in articles in the targeted journals. As is the case with the other strategies, it is unclear whether studies did not use these strategies or whether studies used the strategies but did not describe them in their articles. Regardless, it is difficult for the reader to draw conclusions about the effectiveness of the intervention without this information. For example, without information on provider training, null results could be due to multiple factors: an ineffective intervention, lack of provider skill acquisition, or lack of attention to maintenance of skills over time. Lack of critical detail, such as describing training methods, also poses difficulties for researchers who want to replicate the study or for those who want to translate the treatment to an applied setting.

The strategy that had the lowest percent use among our surveyed articles was the assessment of nonspecific treatment effects, at 6.0%. Nonspecific effects include such factors as the assessment of the quality of provider–patient relationship (e.g., therapeutic alliance) or the assessment of provider variables, such as warmth, empathy, respect for the patient, understanding, trustworthiness, credibility, and knowledge. Measurement should take care not to interfere with the ascertainment of the primary outcome and should also be multifaceted, involving different perspectives (patients, providers), different modalities (self-report, provider ratings, direct observation), and possibly different facets of the individual (cognition, affect, behavior; Kazdin, 1986). These factors are not inert treatment components; they have been shown to increase therapeutic alliance and, in and of themselves, have modest to strong relationships with both treatment retention and outcome (e.g., Klein et al., 2003; Lambert, 1989; Martin, Garske, & Davis, 2000). Without measurement of nonspecific effects, for example, lack of intervention effect cannot be solely attributed to an ineffective intervention; rather it may be due to a provider with weak counseling skills. Measurement of nonspecific effects is particularly critical when different providers are nested within treatment and contact–control groups (Wampold & Serlin, 2000). In that case, without measuring nonspecific effects, an effective

intervention may be due to either the intervention itself or to greater therapeutic alliance between providers and participants in the intervention group. If a significant intervention effect is found, therapeutic alliance should be relatively equal across groups so that treatment outcome differences can be attributed to the active, theory-based, "specific" treatment. There are a number of methodological and statistical factors that must be considered when measuring therapist effects that are beyond the scope of this article, but we refer the reader to several other sources of information (Crits-Christoph & Mintz, 1991; Crits-Christoph, Tu, & Gallop, 2003; Kazdin, 1986; Wampold & Serlin, 2000).

That only 27% of studies assessed whether the intervention was delivered as specified was cause for concern. This percentage increased by only 3% when referenced articles from other journals were included. Thus, it is difficult for readers to determine exactly what was delivered and whether it was faithful to the underlying theory or model. This also has implications for exportability and dissemination. If a successful trial is described but adherence to protocol is not monitored, applications of the study intervention in real world settings may be compromised and/or unsuccessful, potentially at great cost. Some examples of what our coders were looking for in determining whether articles checked adherence to protocol were review of clinician notes, behavioral checklists, supervisor ratings of process notes, audio- or videotaped sessions, and live observation of sessions. Ideally, monitoring adherence to protocol should involve checking both errors of commission (adding in components that were not specified by the protocol) and errors of omission (deleting components that were specified by the protocol) according to an a priori list of criteria. Those coding for protocol adherence should be blind to the treatment condition and try to guess the treatment being administered. The success of the blinding of the subjects should also be reported. Recent meta-analyses showed that only 8% (15/191) of trials (general medicine and psychiatry combined) reported on the success of blinding, and only one third of these reported successful blinding (Fergusson, Glass, Waring, & Shapiro, 2004). Indices of expectancy effects, intervention purity, and cross-contamination could be calculated and used in the analyses.

The mean proportion adherence across all treatment fidelity strategies was .55. This proportion did not significantly change over time but did vary considerably between categories. The category with the highest mean proportion adherence was Design, at .80. The category with the lowest mean proportion adherence was Training at .22. In this latter category, both the median and mode were zero and the mean proportion adherence at the 75th percentile of this category was only .25. A total of 15.5% of articles achieved .80 or greater adherence.

There are several limitations to our study. We identified articles through reviewing every article in the five targeted journals across 10 years. It is conceivable that we may have missed some articles that could have met inclusion criteria.

Also, with the advent of new guidelines (CONSORT, TREND), it may be argued that treatment fidelity reporting will be greatly improved in the future, making our data less relevant. Although we hope that this is true, this remains an empirical question. In this regard, our data can, to some extent, serve as baseline data. CONSORT and TREND, unfortunately, do not include all of the aspects of treatment fidelity included in our framework, and our framework includes items that are specifically relevant for health behavior change researchers. Although treatment fidelity has been on the radar now for more than 30 years, with numerous articles published in prominent journals (e.g., Kazdin, 1986; Lichstein et al., 1994; Moncher & Prinz, 1991; Yeaton & Sechrest, 1981), there continues to be little evidence reported that treatment fidelity is being conducted routinely in clinical trials.

Another limitation of our study is that we cannot ascertain the reason for the low level of treatment fidelity. Reasons could include poor treatment fidelity implementation by the researcher (intentional or unintentional), lack of reporting of treatment fidelity by researchers despite satisfactory implementation, journal editorial policy, space limitations, or culture within a discipline regarding the report of details on treatment fidelity. Although we cannot be sure of the source of the lack of treatment fidelity information, we wanted journal editors and researchers to be aware of the large degree of heterogeneity in reporting treatment fidelity. Our data support the rationale for the need for greater consistency in reporting and also indicate the types of treatment fidelity needed to enable accurate evaluation and replication of studies. Without reporting treatment fidelity, it is impossible for readers to judge whether two treatments were adequately compared, for example, or whether a valid study replication has occurred. Treatment fidelity should be reported to the degree that the reader can be confident that alternative explanations have been ruled out regarding the effect of the independent variable on the dependent variable. An item-by-item accounting of the state of the science provides a starting point to ascertain where improvements are needed.

Finally, it was beyond the scope of this study to analyze whether the articles that demonstrated excellent treatment fidelity also had better outcomes. This is a complicated issue, as one can imagine that there could exist a high degree of treatment fidelity of a poorly conceptualized intervention, and therefore the association between degree of treatment fidelity and outcome would be negative. Thus, treatment fidelity is not always associated with better outcomes. It does, however, provide researchers with confidence that their intervention was adequately tested (even if it was a poorly designed intervention) and that other potential confounders have been ruled out. Nonsignificant results, for example, could be attributed to the possibility that providers were not trained adequately rather than to the type of treatment under investigation. Treatment fidelity prevents the premature rejection of treatments that could be effective as well as the acceptance of treatments that are nonreproducible because of low internal validity. A high degree of treatment fidelity can provide the best test of a well-conceptualized intervention.

With increasing focus by the NIH and other institutions on dissemination of effective treatments, it becomes even more important to monitor treatment fidelity in efficacy trials so the effects have the best chances of being maintained as they become translated into effectiveness trials. Treatment fidelity is important for external validity for two reasons: (a) treatment replication and treatment comparison across studies and (b) evaluation of the treatment in applied settings (Moncher & Prinz, 1991). Henggeler et al. (1997), for example, found that when disseminating their treatment from a clinical to community setting, the intervention effects were attenuated. They attributed this attenuation to the lack of attention to treatment fidelity in the community trial. Specifically, these investigators did not have the intensive provider train-

ing and supervision that they had in their efficacy trial, in order to be more consistent with standard clinical practice and have greater ecological validity. However, Henggeler et al. (1997) have explained that these incremental costs are minimal when compared with the costs of providing services that are ineffective. Although more training and supervision may increase program costs, Dumas, Lynch, Laughlin, Smith, and Prinz (2001) recommended that key components of dissemination projects be regularly checked for adherence to the protocol, perhaps by using procedures that have been streamlined from the efficacy trial. Opponents of treatment fidelity contend that ongoing monitoring of treatment fidelity is costly and not ecologically valid for dissemination research. Dumas et al. (2001) argued that this is a "false opposition," as it is not a matter of choice between

> theoretically and methodologically sound projects that promote high but "unrealistic" levels of intervention fidelity and that are insensitive to community needs, and projects that address these needs but from which little can be learned because they lack sound theoretical and methodologic bases. (p. 46)

Rather, these authors contended that programs should be implemented in such a manner that one can rule out alternative explanations when outcomes are assessed, clearly pointing to what works and what does not work. Future research should develop and validate cost-effective strategies for transferring the rigor of clinic-based efficacy treatment protocols to community settings (Henggeler et al., 1997).

There could be several potential uses of our treatment fidelity assessment tool. It could be a useful guide for researchers who are designing a study or help guide researchers to monitor service delivery and treatment integrity while the trial is ongoing (Dumas et al., 2001). Ongoing process monitoring can prevent experimental drift, ultimately decreasing costs and improving the efficiency and internal validity of the intervention. The assessment tool could also be used to help researchers assess the reasons why their treatment did not work. For example, Project MATCH, a multisite collaborative project designed to evaluate patient–treatment interactions in alcoholism treatment, did not find much empirical support for their a priori matching hypotheses. Carroll et al. (1998) subsequently undertook a rigorous investigation of treatment integrity and discriminability to determine whether threats to treatment fidelity and internal validity could explain the lack of treatment effect. Carroll et al. (1998) found a high degree of treatment integrity, treatment discriminability, and similar levels of therapeutic alliance across treatment conditions, enabling them to rule out these alternative explanations for the modest findings. Our treatment fidelity assessment tool can help guide this process of post hoc evaluation. The assessment tool could also be a checklist for grant reviewers and journal editors evaluating either proposed or completed treatment outcome studies. It could also be a useful teaching tool for students learning about study design and internal and external validity.

We believe that our treatment fidelity framework could also be a useful supplement to both the CONSORT guidelines for randomized controlled trials (RCTs) (Altman et al., 2001; Moher, Schulz, & Altman, for the CONSORT Group, 2001) and the TREND guidelines for reporting of nonrandomized trials (Des Jarlais, Lyles, Crepaz, & the Trend Group, 2004). Our treatment fidelity framework, however, shifts the attention away from more molar design issues to a more molecular examination of the conduct of therapist and client within a given treatment to ascertain if the treatment of interest was given a fair test (Lichstein et al., 1994). Provider Training (skill acquisition and maintenance), Delivery factors (nonspecific effect; protocol adherence), Receipt, and Enactment are either underemphasized or not discussed by the CONSORT guidelines. Davidson et al. (2003) also suggested supplementing the CONSORT criteria by adding behavioral medicine-specific guidelines for reporting behavioral medicine RCTs. Whereas Davidson et al. and the CONSORT criteria focus on general methodology for RCTs (randomization, recruitment, adverse events, etc.), including some aspects of treatment fidelity, our framework hones in on the treatment fidelity aspect and discusses treatment fidelity in more depth by positing five categories of treatment fidelity, specific components within each category, and the specific strategies by which treatment fidelity may be enhanced. We believe that our tool makes an incremental contribution to the CONSORT guidelines because it was developed specifically for health behavior researchers. We hope that this tool can offer health behavior scientists more tailored recommendations on how to improve their scientific reporting, minimize equivocal conclusions regarding treatment efficacy, and improve their ability to translate effective treatments to real-world contexts.

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al., for the CONSORT Group. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663–694.

Bellg, A., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., et al. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the Behavior Change Consortium. *Health Psychology, 23,* 443–451.

Borrelli, B., Resnick, B., Bellg, A., Ogedegbe, G, Sepinwall, D., Orwig, D., & Czajkowski, S. (2002, April). *Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the Behavioral Change Consortium.* Seminar presented at the Annual Meeting of the Society of Behavioral Medicine, Washington, DC.

Burgio, L., Corcoran, M., Lichstein, K., Nichols, L., Czaja, S., Gallagher-Thompson, D., et al. (2001). Judging outcomes in psychosocial interventions for dementia caregivers: The problem of treatment implementation. *The Gerontologist, 4,* 481–489.

Cameron, R., Brown, K. S., Best, J. A., Pelkman, C. L., Madill, C. L., Manske, S. R., & Payne, M. E. (1999). Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. *American Journal of Public Health, 89,* 1827–1831.

Carroll, K. M., Connors, G. J., Cooney, N. L., DiClemente, C. C., Donovan, D. M., Kadden, R. R., et al. (1998). Internal validity of Project MATCH treatments discriminability and integrity. *Journal of Consulting and Clinical Psychology, 66,* 290–303.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapy. *Journal of Consulting and Clinical Psychology, 59,* 20–26.

Crits-Christoph, P., Tu, X., & Gallop, R. (2003). Therapists as fixed versus random effects—some statistical and conceptual issues: A comment on Siemer & Joormann (2003). *Psychological Methods, 8,* 518–523.

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

Des Jarlais, D. C., Lyles, C., & Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94,* 361–366.

Dumas, J. E., Lynch, A. M., Laughlin, J. E., Smith, E. P., & Prinz, R. J. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine, 20*(Suppl. 1), 38–47.

Fals-Stewart, W., Birchler, G. R., & O'Farrell, T. J. (1996). Behavioral couples therapy for male substance-abusing patients: Effects on relationship adjustment and drug-using behavior. *Journal of Consulting and Clinical Psychology, 64,* 959–972.

Fergusson, D., Glass, K. C., Waring, D., & Shapiro, S. (2004). Turning a blind eye: The success of blinding reported in a random sample of randomized placebo controlled trials. *British Medical Journal, 328,* 432–436.

Henggeler, S. W., Melton, G. B., Brondino, M. J., Scherer, D. G., & Hanley, J. H. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. *Journal of Consulting and Clinical Psychology, 65,* 821–833.

Kazdin, A. E. (1986). Comparative outcome studies of psychotherapy: Methodological issues and strategies. *Journal of Consulting and Clinical Psychology, 54,* 95–105.

Klein, D. N., Schwartz, J. E., Santiago, N. J., Vivian, D., Vocisano, C., Castonguay, L. G., et al. (2003). Therapeutic alliance in depression treatment: Controlling for prior change and patient characteristics. *Journal of Consulting and Clinical Psychology, 71,* 997–1006.

Lambert, M. J. (1989). The individual therapist's contribution to psychotherapy process and outcome. *Clinical Psychology Review, 9,* 469–485.

Lichstein, K. L., Riedel, B. W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment implementation model. *Advances in Behavior Research and Therapy, 16,* 1–29.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communications Research, 28,* 587–604.

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68,* 438–450.

Maude-Griffin, P. M., Hohenstein, J. M., Humfleet, G. L., Reilly, P. M., Tusel, D. J., & Hall, S. M. (1998). Superior efficacy of cognitive–behavioral therapy for urban crack cocaine abusers: Main and matching effects. *Journal of Consulting and Clinical Psychology, 66,* 832–837.

Moher, D., Schulz, K. F., & Altman, D. G., for the CONSORT Group. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *JAMA, 285,* 1987–1991.

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

Rotheram-Borus, M. J., Reid, H., & Rosario, M. (1994). Factors mediating changes in sexual HIV risk behaviors among gay and bisexual male adolescents. *American Journal of Public Health, 84,* 1938–1946.

Smith, B., & Sechrest, L. (1991). Treatment of Aptitude × Treatment interactions. *Journal of Consulting and Clinical Psychology, 59,* 233–244.

Stephens, R. S., Roffman, R. A., & Curtin, L. (2000). Comparison of extended versus brief treatments for marijuana use. *Journal of Consulting and Clinical Psychology, 68,* 898–908.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425–433.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156–167.

Received June 21, 2004
Revision received December 28, 2004
Accepted January 3, 2005 ■