Nursing Research • January/February 2007 • Vol 56, No 1, 54–62

Intervention Research

Establishing Fidelity of the Independent Variable in Nursing Clinical Trials

Karen Farchaus Stein ▼ Judy T. Sargent ▼ Nicholas Rafaels

- ***Background:*** Internal validity of a randomized clinical trial of a nursing intervention is dependent on intervention fidelity. Although several methods have been developed, evaluating audio or audiovisual tapes for prescribed and proscribed interventionist behaviors is considered the gold standard test of treatment fidelity. This approach requires development of a psychometrically sound instrument to meaningfully categorize and quantify interventionist behaviors.
- ***Objective:*** To outline critical steps necessary to develop a treatment fidelity instrument.
- ***Methods:*** A comprehensive literature review was conducted to determine procedures used by other researchers. The literature review produced five quantitative studies of treatment fidelity, all in the field of psychotherapy, and two replication studies. A synthesis of methodologies across studies combined with researchers' experiences resulted in identification of the steps necessary to develop a treatment fidelity measure.
- ***Results:*** Seven sequential steps were identified as essential to the development of a valid and reliable measure of treatment fidelity. These steps include (a) identification of the essential elements of the experimental and control treatment modalities; (b) construction of scale items; (c) development of item scaling; (d) identification of the units for coding; (e) item testing and revision; (f) specification of rater qualifications and development of rater training program; and (g) development and completion of pilot testing to test psychometric properties. Development of the Possibilities Project Psychotherapy Coding Questionnaire is described as an illustration of the seven-step process.
- ***Discussion:*** The results show the essential steps that are unique to the development of treatment fidelity measures and show the feasibility of using these steps to construct a psychometrically sound treatment-specific fidelity measure.

- ***Key Words:*** internal validity · intervention fidelity · randomized clinical trials

Randomized clinical trials (RCTs) are vital in advancing effective new nursing interventions. In RCTs, the efficacy of the experimental intervention is established through comparison of patient outcomes between preestablished groups, including one group who received the experimental treatment and one who received the control or comparison treatment. The nature of many experimental nursing interventions is that they are flexible, dynamic, and individualized to be sensitive and responsive to the unique characteristics of the participant. Although this approach is essential to ensure clinically meaningful and relevant interventions, internal validity of the trial is dependent on the systematic and reliable delivery of the independent treatment variable (Calsyn, 2000). Hence, methods to establish reliable delivery of the treatment intervention objectively are central to the integrity of any randomized trial.

Reliable and competent delivery of an experimental treatment by the interventionist is referred to in the literature as *intervention fidelity* (Moncher & Prinz, 1991; Ogrodniczuk & Piper, 1999; Santacroce, Maccarelli, & Grey, 2004). Intervention fidelity has two core components: *adherence* and *competence*. Adherence is the most basic and is the extent to which the interventionists' behaviors conform to the treatment protocol. Adherence is focused on the quantity of prescribed behaviors that are delivered in a treatment session or course, and compares the quantity of generic interventionist behaviors (common across psychotherapy) and behaviors that are proscribed by the protocol. The competence component is more complex and is focused on the interventionist's skillfulness in the delivery of the intervention.

In recent years, important strategies have been used to improve intervention fidelity, including the use of treatment manuals, formal training of interventionists,

Karen Farchaus Stein, PhD, RN, FAAN, is Associate Professor; and Judy T. Sargent, MS, RN, CS, is Doctoral Student and Research Assistant, School of Nursing, University of Michigan, Ann Arbor.

Nicholas Rafaels, MS, is Programmer/Data Analyst, Division of Allergy and Clinical Immunology, Department of Medicine, The Johns Hopkins University, Baltimore, Maryland.

and clinical supervision (Carroll et al., 2000). Evaluating audio or audiovisual tapes for prescribed and proscribed interventionist behaviors is considered the gold standard test of treatment fidelity (Markowitz, Spielman, Scarvalone, & Perry, 2000; Waskow, 1984). Trained and reliable coders who are blind to the treatment approach rate therapist behaviors in a recorded session using an instrument that reflects central elements of the prescribed intervention. It is by far the most difficult method to establish treatment fidelity and has not been addressed generally in nursing treatment trials.

Central to the monitoring process of interventionist behaviors is the availability of a valid and reliable instrument that enables meaningful categorization and quantification of interventionist behaviors. Rating scales must be designed to address the unique content of the interventions used in the study and, at the same time, to demonstrate sound psychometric properties that enable valid and reliable coding of the behavior. To date, little attention has been paid in the methodological literature to the process and issues related to development of instruments to evaluate treatment adherence.

The primary purposes of this paper are to outline critical steps in order to develop an instrument to evaluate treatment fidelity, and to describe the development of the Possibilities Project Psychotherapy Coding Questionnaire (PPPCQ) as an example of a treatment fidelity instrument used in a nursing intervention RCT.

The PPPCQ was developed to establish treatment fidelity in an RCT of a cognitive–behavioral treatment to promote recovery and well-being in women with anorexia nervosa and bulimia nervosa. For this study, the experimental treatment was the Identity Intervention Program (IIP), a 20-week individual psychotherapy that focuses on the development of new positive self-schemas as the means to promote recovery from the eating disorders. The manualized treatment was designed to be delivered by an experienced advanced practice psychiatric mental health nurse. A manualized Supportive Psychotherapy Intervention (SPI) was used as the control treatment.

Treatment Fidelity Measure: Purpose and Uses

Use of a structured instrument enables a systematic quantification of interventionist behaviors delivered in a single session and over a course of treatment, hence, provides a basic measure of the type of treatment delivered (Markowitz et al., 2000). Instruments typically are constructed to tap not only behaviors *prescribed* by the intervention but also behaviors that are *universal* in therapeutic interactions and those *proscribed* by the approach, thus enabling quantification of distinctiveness and purity of the treatment delivered (Markowitz et al., 2000). More specifically, the monitoring approach provides a means to test the hypothesis empirically that the interventionist delivered more behaviors that were consistent with the assigned treatment compared to proscribed or universal behaviors in any single session or across the course of treatment. Also, this structured approach might be used to examine empirically important methodological issues related to consistency in treatment delivery over the course of the trial (e.g., *drift*), as well as interventionist and site effects.

Using a structured instrument also provides a means for exploring factors that influence interventionist behaviors, and a means for assessing the effectiveness of specific behaviors in producing the desired outcomes. Important questions related to the effects of severity and duration of illness on interventionist behaviors (e.g., degree of adherence to the prescribed treatment, reliance of interventions outside of the protocol to maintain alliance) may be investigated when both patient characteristics and interventionist behaviors are measured. Similarly, more refined exploration of the effects of specific treatment components (e.g., homework assignments, standard alliance-building behaviors) on outcomes can be explored systematically.

Several examples of psychometrically sound measures of treatment fidelity are available. For example, Hollon (1984) developed the Collaborative Study Psychotherapy Rating Scale to monitor adherence to three forms of psychotherapy tested in a National Institute of Mental Health multisite depression clinical trial; Markowitz et al. (2000) modified the measure for use with persons with human immunodeficiency virus and depression. Others have developed measures to assess reliable delivery of treatments for substance use disorders (Barber, Mercer, Krakauer, & Calvo, 1996; Carroll et al., 2000). Although each of these measures has been shown to be psychometrically sound and reliable in discriminating treatment approaches, none are appropriate for adherence studies of new forms of treatment that span beyond those specifically addressed by the measures. Also, detailed accounts of the steps required to develop a measure of treatment adherence are not provided in any of these studies.

Developmental Process of the Treatment Fidelity Measure

A comprehensive literature review was conducted to identify the essential steps necessary to develop a measure of treatment fidelity. The following databases were searched: CINAHL, MEDLINE, PsycINFO, and PsycARTICLES. The search terms were: "intervention fidelity"; "treatment fidelity"; "therapist" AND "adherence" AND "competence." The review produced eight quantitative studies of treatment fidelity, all in the field of psychotherapy (Barber & Crits-Christoph, 1996; Barber, Foltz, Crits-Christoph, & Chittams, 2004; Barber et al., 1996; Carroll et al., 2000; DeRubeis, Hollon, Evans, & Bemis, 1982; Hogue, Liddle, Singer, & Leckrone, 2005; Hollon, 1984; Shapiro & Startup, 1992), and two replication studies (Hill, O'Grady, & Elkin, 1992; Markowitz et al., 2000). Three studies were in the area of addictions research (Barber et al., 1996, 2004; Carroll et al., 2000), one was in the area of adolescent problem behaviors (Hogue et al., 2005), and the remainder pertained to the clinical psychological treatment of depression (Barber & Crits-Christoph, 1996; DeRubeis et al., 1982; Hill et al., 1992; Hollon, 1984; Markowitz et al., 2000; Shapiro & Startup, 1992). None of the studies were conducted by nurses, although researchers have suggested that this methodology would benefit nursing research (Santacroce et al., 2004). A synthesis of methodologies combined with our own experiences resulted in identification of a seven-step

process to develop a measure of treatment fidelity. The seven steps include (a) identification of essential elements of the treatment modalities; (b) construction of scale items; (c) development of item scaling; (d) identification of the units for coding; (e) item testing and revision; (f) specification of rater qualifications and development of rater training program; and (g) development and completion of pilot testing to test psychometric properties. Each of the seven steps is described below.

A comprehensive literature review was conducted to identify the essential steps necessary to develop a measure of treatment fidelity.

Step 1: Identification of the Essential Elements of the Treatment Modalities

Definitions and Issues The identification of concrete, specific, and observable interventionist behaviors that are posited to bring about the desired patient change is necessary to differentiate quantitatively between experimental and control treatment types (Carroll et al., 2000). Clients' behaviors and responses are not of interest to fidelity measurement.

To establish adherence, it is essential to identify three categories of interventionist behaviors: *unique*, *common*, and *proscribed* (Calsyn, 2000). First, behaviors that are both essential and specific to the target treatment type must be identified. These behaviors are *unique* factors because they are distinctive to the target treatment and are essential to bring about patient change (Kazdin & Nock, 2003). According to Carroll, Kadden, Donovan, Zweben, and Rounsaville (1994), unique factors are the "active ingredients" (p. 152) in the model essential to its effectiveness in producing emotional or behavioral change (Calsyn, 2000).

The second category includes behaviors *common* across all treatment modalities. A critical determination of whether the "active ingredients" (Carroll et al., 1994, p. 152) of the treatment contributed to observed differences is whether other interventionist behaviors that are essential to positive outcomes are equivalent across treatment types. For example, in most psychosocial treatments, establishing rapport, specifying goals, setting limits, and providing information are viewed as essential interventionist behaviors necessary to promote change. These general (or common) behaviors are viewed as essential to all forms of effective treatment, and therefore, must be held constant to conclude a causal link between the unique interventions and change.

Finally, to draw a causal link between a targeted intervention and outcomes, it is essential that the interventionist is not using strategies from competing interventions. Interventionist behaviors that are unique to the comparative intervention (i.e., control) should not be brought into the target protocol. These are *proscribed* behaviors. Inclusion of proscribed behaviors in the measure provides a means to verify that treatment type blending does not occur.

Identification of the Essential Elements of the Possibilities Project Treatment For this study, it was necessary to identify the essential elements of the IIP (e.g. experimental), the SPI (e.g., control), and basic interventions that are common (common elements; CEs) to all forms of psychotherapy. Written manuals are available for both IIP and SPI treatments, and the essential elements of both therapies were drawn from these documents. Written objectives from the IIP manual are outlined at the start of each unit and provide the framework for the identification of essential therapist behaviors. In addition, the body of each unit was reviewed for outcome goals and related therapist behaviors that were not reflected in the objectives. A total of 20 therapist behaviors *unique* to the IIP were identified from the treatment manual (see Table 1).

The SPI program is based on the proposition that eating disorder symptoms are caused by psychological problems and the lack of coping strategies. The SPI program aims to identify underlying problems and develop more effective coping strategies. The SPI manual unit objectives are global; thus, therapist behaviors were drawn from the detailed descriptions of the unit activities. The SPI elements are considered proscribed behaviors in the IIP condition and essential for the SPI condition (see Table 1).

Identification of the essential elements in each intervention protocol was completed by an iterative process. The process began with the principal investigator (PI) and graduate assistant independently identifying essential elements of each unit in the IIP in consecutive order. Draft lists were compared, discussed, and revised until consensus was reached. The product of this process was brought to the full research team. Members were asked to review individually each unit of the treatment manual and to verify that all content was addressed completely and accurately. Individual work was discussed in group meetings until full consensus was reach. The same process was used to identify the essential elements of the SPI.

Because this trial was designed to test the efficacy of a form of psychotherapy for eating disorders, therapist behaviors that are considered essential to eating disorder therapy and, more generally, to any form of psychotherapy were considered CE. The CE were identified through literature review and discussion with the interventionist and other members of the clinical team. Examples of CE for the PPPCQ are provided in Table 1.

Step 2: Construction of Scale Items

Definitions and Issues Unique, common, and proscribed elements identified in the previous step are translated into individual statements of observable interventionist behaviors. The level of complexity in the prescribed and proscribed protocols will determine the level of difficulty associated with writing scale items. The language used to construct items must reflect unique and objective behaviors that may be recognized reliably by independent raters. Each item must address a distinct behavior such that overlap and confusion between items does not occur.

TABLE 1. Examples of Objectives, Elements, and Items from the Possibilities Project Psychotherapy Coding Questionnaire

Objective	Element	Item
A. Identity Intervention Program		
Overview of the IIP	Provide an overview of the IIP	The therapist introduced the overall aims of the IIP.
		The therapist provided a description of the theory underlying the IIP.
Introduction to basic concept of identity	Introduction to the basic concept of identity	The therapist introduced the client to the basic concept of identity.
Introduction to the basic concept of "possible self"	Introduction to the concept of "possible self"	The therapist introduced the client to the concept of a "possible self."
Select a desired "possible self" to focus on as a first goal	Select a desired "possible self" to focus on as a goal	The therapist assisted the client to identify (or select) a desired "possible self" (singular) to focus on as a goal.
B. Supportive Psychotherapy Intervention		
Underlying issues are identified	Identify up to four underlying issues	The therapist identifies up to four underlying problems for the client.
Expression of affect is encouraged.	Encourage the *expression* of feelings, thoughts, and opinions	The therapist encourages the client's expression of her own ideas, feelings, and thoughts.
The patient is helped to identify her feelings and thoughts.	Identify feelings, thoughts, and opinions	The therapist assisted the client to *identify* her own ideas, feelings, and opinions.
Encourage assertiveness	Encourage assertiveness	The therapist encouraged the client to identify and use assertiveness skills.
C. Common Elements		
Establish mutual expectations	Establish mutual expectations	The therapist worked with the client to establish mutual expectations from the therapeutic process.
		The therapist discussed basic rules/parameters of therapy (procedure for cancellation of appointments, ground rules, etc…).
Maintain client safety	Maintain client safety	The therapist stressed to the client the importance of staying both emotionally and physically safe.
		The therapist assessed suicidality risk.
Assess the status of current eating disordered symptoms	Assess the status of current eating disordered symptoms	The therapist "checked-in" with the client to explore the status of current eating disordered symptoms.
Obtain general psychiatric history	Obtain general psychiatric history	The therapist obtained a psychiatric history from the client (not specifically focused on the eating disorder).

Note. IIP = Identity Intervention Program; SPI = Supportive Psychotherapy Intervention.

Construction of Items for the PPPCQ scales Unique elements of the IIP and SPI protocols were translated into statements that reflect distinct and observable nurse therapist behaviors (see Table 1). A comprehensive list of unique IIP elements, organized by unit, was used to develop the IIP scale items. Each unique element was translated into the form of a nurse therapist's verbal behavior; attention was paid to each verb to ensure that it connoted a discrete, observable, and consistent set of verbal behaviors. A list of verbs used to describe therapist behaviors and discrete definitions was developed and included as part of the coding instruction.

Step 3: Development of Item Scaling

Definitions and Issues Once items have been constructed, they must be scaled. Scaling options range from a simple dichotomy that reflects the occurrence or nonoccurrence of a behavior to a Likert scale to capture intensity of the behavior. In a study by Carroll et al. (2000), therapist behaviors were coded on a Likert-type continuum of 1 (the behavior is *not at all* present) to 5 (the behavior is *extensively* present). Similarly, Barber et al. (1996) rated the adherence from 1 (*low*) to 7 (*high*). Several disadvantages of the more refined Likert scale have been cited. First, the Likert model may be incongruent with the assumptions of the underlying intervention model and, therefore, be inappropriate for use in many types of intervention trials (Ogrodniczuk & Piper, 1999; Waltz, Addis, Koerner, & Jacobson, 1993). Intervention approaches very rarely suggest that the interventionist strives to deliver a target behavior as many times as possible during the session

(Waltz et al., 1993). Yet, a fidelity measure based on a Likert scale assumes that higher levels or frequencies of the behaviors are desired. A second disadvantage is that a Likert-type scale introduces a greater level of subjectivity into the ratings reducing the robustness of the model and, therefore, is likely to influence negatively attempts to establish interrater reliability (Waltz et al.). Finally, Likert scaling requires more sophisticated raters and takes more time to train, reach, and maintain acceptable interrater reliability. Therefore, it is likely to be more resource intensive.

In contrast, dichotomous rating simply notes whether an intervention behavior occurred or not in a session, and, when summed across a session or a series of sessions, provides a count of intervention behaviors within a category (e.g., unique or common behaviors). The two disadvantages are: (a) It does not take into account the amount of time invested by the interventionist in a target behavior and risks providing a distorted picture of the distribution of interventionist focus in a session; and (b) It may have a greater standard error compared to the Likert model; this increase may hinder the ability to detect differences in coder scores, and thus, to detect lack of interrater reliability. However, the dichotomous model eliminates outliers and clustering, thus improving robustness and giving a more honest assessment of differences in coder scores and interrater reliability. Dichotomous scaling enables more clearly specified items for coding and greater interrater reliability at lower cost (Waltz et al., 1993).

Item Scaling for the PPPCQ Scales Dichotomous (*yes* = 1, *no* = 2) scaling was selected for the PPPCQ in an effort to increase interrater reliability and decrease subjective interpretation of the individual items. Each item of the IIP, SPI, and CE scales is coded simply as *present* = 1 or *absent* = 0. The dichotomous scale was chosen after the group decided that the models underlying the intervention protocols were not based on a quantification of frequency or intensity of therapist behaviors.

Step 4: Identification of the Units for Coding

Definitions and Issues An important decision in developing a measure of treatment fidelity is to define what will be treated as a codeable unit. Three distinct definitions of codeable units were identified:

a. In "event-by-event coding" (Waltz et al., 1993, p. 622), a single unit is considered a therapist's turn (e.g., from start of interventionist utterance to start of participant utterance; for an example, see Wills, Faitler, & Snyder, 1987). The unit is the interventionist behavior(s) that occurred during that turn. This approach is likely to lead to many uncodeable units because not all of the interventionist's utterances reflect a complete intervention behavior. In fact, any single intervention is likely to occur over several turns, making coding of each one difficult, and introducing the risk that a single intervention will be coded multiple times. However, this approach has the advantage of providing a measure of the percentage of interventionist turns dedicated to a specific intervention. Furthermore, because each interventionist's turn has a code, this approach enhances the coder's ability to identify specific areas of disagreement.
b. The "occurrence–nonoccurrence" (Waltz et al., 1993, p. 623) approach focuses on the total collection of therapist utterances during a session as the codeable unit. This approach addresses the question of whether a specific intervention behavior occurred in the session (e.g., see Ogrodniczuk & Piper, 1999; Shapiro & Startup, 1992). The advantages and disadvantages are directly opposite from those cited for the "event-by-event" coding. This coding reflects an overview of the interventionist behaviors that occurred within a session, and avoids issues related to coding a behavior when it occurs over the period of several interventionist turns.
c. A final approach found in the literature is focused on randomly selected *timed segments* of a treatment session (for an example, see Luborsky & DeRubeis, 1984). Like the occurrence–nonoccurrence approach, this approach focuses on an overall collection of interventionist turns; however, it is focused only on a predetermined segment of the session. Luborsky and DeRubeis were the only investigators found who used this approach, and the method used to identify the 15-minute coding segments was not described. The sole advantage to this type of approach appears to be the reduced effort and cost associated with the limit proportion of the session coded.

Identification of the Unit for Coding for the PPPCQ Scales The unit selected for coding for the measure was the "occurrence–nonoccurrence" (Waltz et al., 1993, p. 623) approach. The coders reviewed a session in its entirety and determined whether an interventionist behavior occurred.

Step 5: Item Testing and Revision

Definitions and Issues To refine definitions of key behaviors and coding instructions, item clarity and specificity must be established. Experts (e.g., PI, key personnel) with the population of interest and in the intervention approach code a subset of intervention sessions. Individual and group coding sessions, along with group discussion of coding decisions, are used to clarify and revise items and instructions until an acceptable level of consensus is reached (Barber et al., 1996). This process may result in several iterations of the measure until the group agrees that the measure is acceptable and ready for pilot testing.

Item Testing and Revision for the Questionnaire A two-step iterative process was used to test and revise items in the three scales that comprise the PPPCQ. Using case sessions from a pilot study of the full clinical trial, the PI and the graduate student assistant coded sessions individually and then discussed coding outcomes. Items were revised until both agreed that group input was warranted. In the second phase, treatment sessions from the pilot study were coded individually by all members of the research

team, and group discussions were used to compare and discuss codings and revise scale items. Group consensus was used to determine the point at which the PPPCQ was ready for pilot testing. The first version contained 64 items. There were eight iterations before the final version, which contained 98 items.

Step 6: Specification of Rater Qualifications and Development of Rater Training Program

Definitions and Issues The issue of target respondent is critical. In this case, the respondent is the rater who will use the measure to code the content of the intervention session. In most fidelity studies reviewed, the explicitly stated assumption was that coding of intervention sessions is a complex task that requires expertise comparable to that of the study interventionist. Expertise was specified in terms of disciplinary background, level of training, and years of clinical experience (Barber et al., 1996; DeRubeis et al., 1982; Shapiro & Startup, 1992). Clinical experience was defined in terms of the population of interest and the treatment technique and approach (Carroll et al., 2000). In a minority of studies, graduate students in clinical programs were used as coders (Hill et al., 1992). In only one study were raters with no expertise in the field used (Ogrodniczuk & Piper, 1999).

Training of the coders to ensure reliable utilization of the measure is critical and is defined typically in terms of duration and methods. Most investigators describe coder training as comparable in duration and intensity to that of the interventionist. Methods include didactic training, independent review of the treatment manual (DeRubeis et al., 1982), and group coding meetings to discuss the intervention elements and meaning of coding items and practices (Carroll et al., 2000; Hill et al., 1992). Practice codings include both group and individual coding sessions; results are then discussed in a group setting. Researchers have developed and used a rater's manual that included both general directions for coding and detailed descriptions and examples of the behaviors associated with each item (Carroll et al.). In a number of studies, individual coding was evaluated against expert consensus coding to establish requisite initial interrater reliability. Efforts to maintain high interrater reliability included periodic recalibration codings and intermittent reliability assessments (Carroll et al.).

Rater Qualifications and Training for the Questionnaire

Given that the IIP and SPI protocols were designed to be delivered by an advanced level practitioner in psychiatric mental health nursing, it was decided that the independent raters must possess similar qualifications. The rater training procedures were similar to those used to train the nurses administering the experimental and control interventions. The raters were oriented to the intervention protocols through didactic training (e.g., presentation by the PI focusing on the theoretical framework and empirical groundings of the interventions), detailed discussion of the intervention strategies and aims, case presentations, and clinical discussions. Next, the raters were given the opportunity to code practice audiotapes, which were evaluated with respect to the expert consensus ratings of the same tapes. Raters were *certified* to code tapes independently. Interrater reliability was established at 93.4% before independent coding for the pilot study was initiated.

Step 7: Development and Completion of Pilot Testing to Test Psychometric Properties

Definitions and Issues The effectiveness of adherence monitoring in establishing treatment fidelity is dependent on the psychometric soundness of the treatment adherence instrument. To establish validity and reliability of the measure, a pilot study is necessary. Discriminant validity can be addressed by comparison of the study groups on the scale scores. Thus, it would be expected that the experimental group would score higher on the experimental treatment scale compared to the control group, and vice versa. However, no differences between experimental and control groups would be expected on the CE scale.

Confirmatory factor analysis has been used also to examine the underlying conceptual structure of the measure and its consistency with the theoretical model. Samples from the literature that used this approach of validity assessment ranged from 48 rated sessions (12 tapes rated by 4 coders; DeRubeis et al., 1982), to 156 sessions (12 cases rated by 13 coders; Carroll, Conners, et al., 1998), to 720 sessions (4 sessions of each of 180 patients; Hill et al., 1992). In all cases, the observations used in these analyses were not independent. Ignoring this lack of independence can affect the precision of the estimate, and the underlying structure of the measure may be in question. Independent observations required either one session per case in the analyses or a separate rater for each session. Increasing the number of raters would, in turn, increase the necessary sample size, which could not be attained. Typically, there should be at least 10 cases for each item in the instrument being used (Garson, 2006).

Internal consistency reliability is concerned with the homogeneity of items comprising a scale (DeVellis, 1991). A scale is considered internally consistent to the extent that its items are highly intercorrelated. High inter-item correlation suggests that all items are measuring the same construct. Cronbach's alpha is used widely as a measure of internal consistency. Cronbach's alpha coefficients should be calculated for each conceptually distinct scale in a measure.

Interrater reliability is assessed typically using the intraclass correlation coefficient (ICC) to compare scores of two or more coders who have independently rated a set of intervention sessions. The ICCs are computed for each scale (e.g., unique, common, and proscribed interventionist behaviors) and reflect the proportion of total variance accounted for by variation within taped sessions. Different models for ICC are determined by the nature of how coders are selected, where main effects can be coders, tape sessions, or both. In one scenario, each tape session can be rated by a different set of k coders, selected from a larger population of coders. The resulting model would have a one-way random effects design, as effects due to coders, tape sessions, and random error are not separable. One could also select k coders from a larger population, in

which each coder rates each tape session. Each coder's effect can now be estimated separately and a two-way random effects model can be used. In this design, the coders are the only ones of interest, and each tape session is rated by each coder. This design is similar to the second except that the coder effect is fixed, which results in a two-way mixed model (Shrout & Fleiss, 1979).

Using ICC is preferred over using the Kappa statistic to assess overall scale interrater reliability because Kappa determines agreement by item. If the number of items is increased, the Kappa statistic tends to underestimate the level of agreement, and ICC is less sensitive to these changes (Maclure & Willett, 1987). However, the Kappa statistic may be beneficial at the early stages to evaluate and identify individual specific problematic scale items for rater agreement.

Development and Completion of Pilot Testing to Test Psychometric Properties The sample included seven women aged 18–36 years with a *DSM-IV* diagnosis of anorexia or bulimia nervosa enrolled in a pilot study of a new experimental nurse therapy (IIP) compared to a control condition (SPI). Three subjects were assigned to the IIP condition and four to the SPI condition. A total of 15 audiotaped and transcribed psychotherapy sessions (a random selection of audiotapes) were rated by two independent, qualified coders. The coders were experienced psychiatric mental health nurses, one a second-year graduate student in psychiatric mental health nursing practitioner program and the other a master's-prepared nurse with clinical specialist certification in adult psychiatric mental health nursing. Coders completed an extensive training program that included didactic training, protocol manual orientation, theoretical overview provided by the primary investigator, practice individual and group codings of audiotapes, weekly meetings with the primary investigator to clarify inconsistencies and develop consensus on the meaning of individual questionnaire items, and review of the practice codings until an acceptable level of interrater reliability was established at 93.4%. Periodic reevaluation was conducted to ensure reliability and to prevent the possibility of rater drift.

To assess discriminant validity, means were computed for the PPPCQ subscales by group. The mean proportion of scale items (e.g., interventionist behaviors) present in a session by group for Raters 1 and 2 is shown in Table 2. As predicted, individuals in the IIP group were higher on the IIP subscale compared to individuals in the SPI group. Individuals in the SPI group scored higher on the SPI subscale when compared to individuals in the IIP group, providing evidence of construct validity. Within treatment group, individuals in the SPI group scored significantly higher on the SPI subscale than on the IIP or CE subscales. However, for individuals in the IIP group, there was no significant difference between scores in the IIP and SPI subscales. An examination of mean scores shows that the interventionist used a relatively small proportion of IIP strategies relative to SPI strategies. This difference is due to the fact that the IIP interventions are written at a molar level, focusing on the core themes of a session, such that only a few strategies would be used in a single session. In contrast, the SPI interventions address more molecular behaviors designed to support and enhance participant self-exploration.

Internal consistency of the three subscales of the PPPCQ was assessed using Cronbach's alpha coefficients. Cronbach's alpha coefficients for the PPPCQ subscales were acceptable; IIP = .79, SPI = .84, CE = .45. The CE scale included a diverse collection of behavior, including orientation and basic ground rules of therapy, monitoring of ED behaviors and health status, and termination-related therapist behaviors. As such, a high alpha coefficient would not be expected.

Interrater reliability was addressed in two phases. First, Kappa coefficients, with a theoretical range from 0 to 1, were computed for each item. Ideally, it is desirable

TABLE 2. Possibilities Project Psychotherapy Coding Questionnaire Subscale Means (%)

	IIP Group	SPI Group
Group Rater 1		
IIP Subscale	M = 10.92 (SD = 0.31)	M = 0.29 (SD = 0.05)*
SPI Subscale	M = 6.79 (SD = 0.25)	M = 36.11 (SD = 0.48)*
CE Subscale	M = 15.15 (SD = 0.36)†	M = 27.27 (SD = 0.45)*‡
Group Rater 2		
IIP Subscale	M = 13.41 (SD = 0.34)	M = 0.57 (SD = 0.08)*
SPI Subscale	M = 11.11 (SD = 0.32)	M = 42.59 (SD = 0.50)*
CE Subscale	M = 19.19 (SD = 0.39)	M = 22.72 (SD = 0.42)*§

Note. IIP = Identity Intervention Program; SPI = Supportive Psychotherapy Intervention; CE = Common Elements.
*Significantly different ($p < .05$) for group comparison across subscale.
†Multiple comparison Games–Howell test demonstrated a significant difference ($p < .05$) between the CE and SPI subscales within the IIP group.
‡Multiple comparison Games–Howell test demonstrated significant differences ($p < .05$) between IIP and SPI subscales and between IIP and CE subscales within the SPI group.
§Multiple comparison Games–Howell test demonstrated a significant differences ($p < .05$) among each of the subscales for the SPI group.

to have high Kappa scores. However, the Kappa statistic for items with perfect agreement is paradoxically low because of lack of variance (Feinstein & Cicchetti, 1990). For the PPPCQ, item Kappas ranged from .60 to 1.0 (once paradoxically low values were deleted), and they were used by the research team to identify and address specific coding problems. Second, ICCs were calculated to provide an estimate of interrater reliabilities, using a sample of 15 randomly selected tapes that were rated by the two raters (n = 30). Using the Shrout and Fleiss (1979) random effects model to estimate reliabilities for independent samples, the three scales were highly reliable. The ICCs were .97 for the IIP scale (58 items), .96 for the SPI scale (18 items), and .78 for the CE scale (22 items).

Development of an instrument to measure treatment fidelity requires systematic completion of a series of seven steps.

The pilot testing provides evidence to support discriminant validity, interrater reliability, and internal consistency of the scales used. Additional testing is ongoing as part of the larger study.

Discussion

Development of an instrument to measure treatment fidelity requires systematic completion of a series of seven steps beginning with identification of essential elements of the intervention protocols and concluding with a pilot study to establish validity and reliability of the measure. Once developed, this instrument can be used to establish treatment fidelity and internal validity of the clinical trial, and to investigate relationships among specific components, participant characteristics, and intervention outcomes.

The development of the PPPCQ provides evidence to support the feasibility of the seven-step approach to treatment fidelity instrument development. It was possible to move systematically from utilization of treatment manuals to identify active and proscribed treatment ingredients and more basic literature to identify the common elements. Independent coding and team collaboration were central to further development and refinement of the measure. Finally, pilot testing was used successfully to evaluate the psychometric properties of the measure. The results of the pilot study provided evidence of the feasibility of utilizing the PPPCQ to evaluate treatment fidelity and to support the psychometric soundness of the measure. Based on these results, a decision was made that no additional modifications of the PPPCQ were needed at this time and a treatment fidelity evaluation for the full RCT was initiated.

The challenges associated with establishing intervention fidelity in an RCT obviously extend beyond the development of a valid and reliable instrument. Overall, a fidelity study requires a high level of commitment of time, energy, and financial resources. All treatment sessions must be recorded and the quality of the recordings must be monitored carefully and consistently to ensure that the population of taped sessions is complete, hence, free of bias. Because resources limit the number of sessions that can be transcribed and coded, a sample size (number of sessions) and sampling plan must be developed that are feasible, are sufficiently powered, and address the issue of interventionist consistency across time within an individual case (e.g., from first to last session for a single case) and across cases (from first to last case randomized to treatment). In addition, other measures, such as an interventionist assessment of strategies used in a session (Carroll, Nich, & Rounsaville, 1998), a participant assessment of interventions received in a session (Gaston & Marmar, 1994), and a measure of competence in delivery of the interventions (Rounsaville, O'Malley, Foley, & Weissman, 1988; Waltz et al., 1993), may be developed and used to broaden the scope of the fidelity assessment and to examine convergence among varying perspectives (e.g., the blind rater, interventionist, and participant).

The credibility of nursing research will be enhanced when researchers provide evidence of fidelity of the independent (treatment) variable. The seven-step process detailed above is one important method by which nursing researchers can achieve this goal. Standardization of psychosocial nursing interventions will help ensure validity of these studies, improve consistency, and enhance comparability across research findings. ◼

Accepted for publication October 4, 2006.

Karen Stein was supported by National Institutes of Health, National Institute of Nursing Research (NINR) Grants R01 05277-01, 1 R55 NR 05277-01, and Judy Sargent was supported by an NIH Institutional NRSA predoctoral traineeship during the preparation of this article. Thank you for the contributions of Nora Arato, Lucy Miller, Adam Lewis, and Amelia Deschamps, members of the Possibilities Project Research Team, in the preparation of this manuscript.

Corresponding author: Karen Farchaus Stein, PhD, RN, FAAN, School of Nursing, University of Michigan, 400 North Ingalls, Ann Arbor, MI 48109 (e-mail: kfarchau@umich.edu).

References

Barber, J., & Crits-Christoph, P. (1996). Development of a therapist adherence/competence rating scale for supportive-expressive dynamic psychotherapy: A preliminary report. *Psychotherapy Research, 6*, 81–94.

Barber, J. P., Foltz, C., Crits-Christoph, P., & Chittams, J. (2004). Therapist's adherence and competence and treatment discrimination in the NIDA Collaborative Cocaine Treatment Study. *Journal of Clinical Psychology, 60*, 29–41.

Barber, J. P., Mercer, D., Krakauer, I., & Calvo, N. (1996). Development of an adherence/competence rating scale for individual drug counseling. *Drug and Alcohol Dependence, 43*, 125–132.

Calsyn, R. J. (2000). A checklist for critiquing treatment fidelity studies. *Mental Health Services Research, 2*(2), 107–113.

Carroll, K., Conners, G., Cooney, N., DiClemente, C., Donovan, D., Kadden, R., et al. (1998). Internal validity of project MATCH treatments: Discriminability and integrity. *Journal of Consulting and Clinical Psychology, 66*(2), 290–303.

Carroll, K., Kadden, R., Donovan, D., Zweben, A., & Rounsaville, B. (1994). Implementing treatment and protecting the validity of the independent variable in treatment matching studies. *Journal of Studies on Alcohol, 12*(Suppl.), 149–155.

Carroll, K., Nich, C., & Rounsaville, B. (1998). Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. *Psychotherapy Research, 8*(3), 307–320.

Carroll, K., Nich, C., Sifry, R., Nuro, K., Frankforter, T., Ball, S., et al. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence, 57*, 225–238.

DeRubeis, R., Hollon, S., Evans, M., & Bemis, K. (1982). Can psychotherapies for depression be discriminated? A systematic investigation of cognitive therapy in interpersonal therapy. *Journal of Consulting and Clinical Psychology, 50*(5), 744–756.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage Publications.

Feinstein, A., & Cicchetti, D. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543–549.

Garson, C. D. (2006). *Factor analysis*. Retrieved April 29, 2006, from http://www2.chass.ncsu.edu/garson/pa765/factor.htm.

Gaston, L., & Marmar, C. (1994). The California psychotherapy alliance scales. In A. Horvath & L. Greenberg (Eds.). *The working alliance: Theory, research, and practice*. Oxford, England: Wiley.

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the collaborative study psychotherapy rating scale to rate therapist adherence in cognitive behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology, 60*(1), 73–79.

Hogue, A., Liddle, H., Singer, A., & Leckrone, J. (2005). Intervention fidelity in family-based prevention counseling for adolescent problem behaviors. *Journal of Community Psychology, 33*, 191–211.

Hollon, S. D. (1984). Final report: System for rating psychotherapy audiotapes. Rockville, MD: Department of Health and Human Services.

Kazdin, A. E., & Nock, M. K. (2003). Delineating mechanisms of change in child and adolescent therapy: Methodological issues and research recommendations. *Journal of Child Psychology and Psychiatry, 44*(8), 1116–1129.

Luborsky, L., & DeRubeis, R. J. (1984). The use of psychotherapy treatment manuals: A small revolution in psychotherapy research. *Clinical Psychology Review, 4*(1), 5–14.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology, 126*, 161–169.

Markowitz, J. C., Spielman, L. A., Scarvalone, P. A., & Perry, S. W. (2000). Psychotherapy adherence of therapists treating HIV-positive patients with depressive symptoms. *Journal of Psychotherapy Practice and Research, 9*, 75–80.

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

Ogrodniczuk, J. S., & Piper, W. E. (1999). Measuring therapist technique in psychodynamic psychotherapies: Development and use of a new scale. *Journal of Psychotherapy Practice and Research, 8*(2), 142–154.

Rounsaville, B., O'Malley, S., Foley, S., & Weissman, M. (1988). Role of manual-guided training in the conduct and efficacy of interpersonal psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 56*(5), 681–688.

Santacroce, S. J., Maccarelli, L. M., & Grey, M. (2004). Intervention fidelity. *Nursing Research, 53*(1), 63–66.

Shapiro, D. A., & Startup, M. (1992). Measuring therapist adherence in exploratory psychotherapy. *Psychotherapy Research, 2*(3), 193–203.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–429.

Waltz, J., Addis, M., Koerner, K., & Jacobson, N. (1993). Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620–630.

Waskow, I. E. (1984). Specification of the technique variable in the NIMH Treatment of Depression Collaborative Research Program. In J. B. W. Williams & R. L. Spitzer (Eds.), *Psychotherapy research: Where are we and where should we go?* New York: Guilford Press.

Wills, R. M., Faitler, S. L., & Snyder, D. K. (1987). Distinctiveness of behavioral versus insight-oriented marital therapy: An empirical analysis. *Journal of Consulting and Clinical Psychology, 55*(5), 685–690.