

In: P.C. Kendal, J.N. Butcher, & G.N. Holmbeck (Eds.), *Handbook of Research Methods in Clinical Psychology*, 2<sup>nd</sup> Ed. New York: Wiley, 1999.

## Chapter 15

---

# **TREATMENT PROCESS RESEARCH METHODS**

WILLIAM B. STILES, PH.D., LARA HONOS-WEBB, M.A., and LYNNE M. KNOBLOCH, M.A.

What happens in psychotherapy? How do therapies differ? How do clients act and think differently? What are the common factors across different therapies? Which are the effective ingredients? What happens as clients improve? We have organized this chapter around this series of questions and we have used illustrative studies to focus on the methods psychotherapy researchers have used to address them.

The questions are interesting because of the answers to two prior questions. The first question people usually ask about psychotherapy is, Does it work? The consensus answer to this question is now clearly Yes—usually (e.g., Kendall & Chambless, 1998; Lambert & Bergin, 1994; Lipsey & Wilson, 1993; Seligman, 1995; M. L. Smith, Glass, & Miller, 1980). Methods of addressing this question are described in other chapters in this volume (e.g., Kendall, Flannery-Schroeder, & Ford, this volume). The broadly positive answer is crucial for process research because it suggests that what happens in therapy is worth studying.

The second prior question arises because there are so many psychotherapies. Approaches to adult individual psychotherapy include psychoanalytic, psychodynamic, cognitive, behavioral, interpersonal, client-centered, gestalt, archetypal, personal construct, process-experiential, reality, and solution-focused, each of which has multiple subtypes. There are also many varieties of child, couples, family, and group therapies. Counts of alternatives run into the hundreds (Herink, 1980; Kazdin, 1986). The differences are not merely technical; the theories underlying the alternatives differ in their understanding of the nature of personality, psychopathology, client change, preferred intervention strategies, and the scope, length, and depth of the therapeutic enterprise.

The second prior question is, then, Which psychotherapy is best? The surprising, though controversial, answer is the Dodo verdict, from *Alice's Adventures in Wonderland*: "Everybody has won, and all must have prizes" (Carroll, 1865/1946, p. 28, italics in original; first quoted in this context by Rozenzweig, 1936, p. 412, and frequently repeated by reviewers and critics, e.g., Beutler, 1991, p. 165; Frank, 1973, p. 1; Grencavage & Norcross, 1990, p. 372; Luborsky, Singer, & Luborsky, 1975, p. 995; Stiles, Shapiro, & Elliott, 1986, p. 165; Wampold et al., 1997). Reviews of outcome research (e.g., Lambert & Bergin, 1994; Lipsey & Wilson, 1993; Luborsky et al., 1975; Smith et al., 1980; Wampold et al., 1997, p. 203) and major comparisons

of contrasting psychotherapeutic approaches (e.g., Elkin, 1994; Elkin et al., 1989; Greenberg & Watson, 1998; Shapiro et al., 1994; Strupp & Hadley, 1979) have reported that most therapies tend to yield more or less equivalent positive changes in clients. This widely replicated result has been described as the *equivalence paradox*—the equivalence of outcomes despite the apparent nonequivalence of theory and the treatment process (Stiles et al., 1986).

Of course, no two psychological procedures have exactly equivalent effects (the null hypothesis is never really true; Meehl, 1978), the degree of equivalence of outcomes remains controversial (e.g., Beutler, 1991; Crits-Christoph, 1997; Howard, Krause, Saunders, & Kopta, 1997; Norcross, 1995), and there may be exceptions (e.g., Emmelkamp, 1994, made a strong case for the superiority of in vivo exposure methods over other behavioral techniques for treating phobias). Nevertheless, the failure to yield differences in effectiveness of a magnitude comparable to the theoretical differences among treatments has been a continuing puzzle.

Process researchers have responded vigorously to the equivalence paradox. Investigators have asked whether the therapies really differ as much as the theories suggest: If the differences were only theoretical, the outcome equivalence would not be paradoxical. They have asked whether there are systematic differences among clients: Interactions with client differences might obscure differences in treatment effectiveness. They have asked what factors are common across apparently different therapies: The outcome equivalence might reflect common effective ingredients that override superficial differences. In this chapter, we consider how investigators have addressed each of these questions. We begin, however, with a more descriptive treatment process question.

### WHAT HAPPENS IN PSYCHOTHERAPY?

Logically and historically, treatment process research begins with naming, describing, classifying, and counting what therapists and clients do. That is, researchers must begin by developing measures. They have done this with great energy. One of the most salient features of treatment process research is the profusion of measures. Researchers have developed thousands of categories and scales, and they have organized these into hundreds of measuring instruments and systems of classification. We have space to mention only a few of them as illustrations in this chapter. More extensive descriptions of some of the better-constructed process instruments and systems have been collected in volumes edited by Kiesler (1973), Greenberg and Pinsof (1986b), and A. P. Beck and Lewis (in press).

So many systems of process classification have been developed that there is even a literature on *meta-classification*—that is, classification of classifications (Elliott, 1991; Elliott & Anderson, 1994; Greenberg, 1986; Greenberg & Pinsof, 1986a; Lambert & Hill, 1994; Russell, 1988; Russell & Staszewski, 1988; Russell & Stiles, 1979). To convey an idea of the variety of ways treatment process has been described, we list some meta-classificatory principles, ways in which process categories and measures differ:

1. *Size of the scoring unit* (e.g., single words or gestures; phrases, clauses, sentences; speaking turns; topic episodes; timed intervals of various durations,

whole sessions; phases of treatment, whole treatments, series of treatments). Measures that target whole sessions are sometimes described as *session impact* measures (Stiles, 1980); measures that target whole treatments may be considered outcome measures. Kiesler (1973) distinguished the *scoring unit* (the material to which the measure is directly applied) from the *contextual unit* (the material that coders or raters are told to consider when assigning the score, which may be considerably larger) and from the *summarizing unit* (the material over which scores are aggregated). For example, for the category "interpretation," the scoring unit might be a single sentence, the contextual unit might include the preceding speech (or some larger context), and the summarizing unit might be the session (or some segment of the session), in which a percentage of interpretations is calculated.

2. *Perspective*. Whose viewpoint is used (therapist, client, external observers, or judges)? Perhaps not surprisingly, the process often looks different from the inside than from the outside and from the perspective of clients than of therapists (e.g., Dill-Standiford, Stiles, & Rorer, 1988).
3. *Data format and access strategy*. What materials from the treatment are studied (e.g., transcripts, session notes, audiotape, videotape, current experience, post-session recall, long-term recall)? How are the materials observed (e.g., observation, self-report, tape-assisted recall)? Tape-assisted recall is a procedure in which audio- or videotape recordings of therapy sessions are replayed for participants, who code, rate, or describe the experience they were having at the time of the recording (e.g., Elliott, 1986; Kagan, 1975).
4. *Measure format* (e.g., coding, rating, verbal description, Q-sort, questionnaire). Coding refers to classifications into nominal categories. Rating refers to placement on an (at least) ordinal scale. Q-sort refers to a procedure in which descriptors are sorted according to how well they characterize the target (e.g., Ablon & Jones, 1998; Jones & Pulos, 1993).
5. *Level of inference*. Marsden (1971), following Berelson (1952), distinguished the *classical strategy*, in which only observable behavior is coded or rated by judges, from the *pragmatic strategy*, in which the coders or raters make inferences about the speaker's thoughts, feelings, intentions, or motivations based on the observed behavior. The classical strategy generally yields higher reliability, but pragmatic schemes may allow raters to integrate and weigh clinically relevant information not encompassed by classical categories.
6. *Theoretical orientation* (e.g., psychoanalytic, experiential, cognitive, behavioral, interpersonal). Some measures aim to assess therapy within particular schools, whereas others claim broader applicability.
7. *Treatment modality* (e.g., individual adult, child, family, group therapy).
8. *Target person(s)*: the focus of the measurement (therapist, client, dyad, family, group).
9. *Communication channel* (e.g., verbal, paralinguistic, kinesic).
10. *Aspect/attribute/feature/dimension*. Among the verbal coding measures, *content categories*, which deal with semantic meaning, have been distinguished from

*speech act categories*, which concern what is done when someone says something. For example, the utterance "What about the situation made you frightened?" might be coded as concerning the content "fear" but the speech act "question." *Paralinguistic measures* concern behaviors that are not verbal but accompany speech (hesitations, dysfluencies, emphasis, tonal qualities). Evaluative ratings (which require some judgment of quality or competence) can be distinguished from descriptive ratings.

Why are there so many measures? In 1973, Kiesler complained that many process measures were developed and then never used again, with many reinvented by researchers who were uninformed of the previous work. Researchers' awareness of previous work now seems to have improved; however, the proliferation of measures has continued. Using standard instruments would offer the advantages of continuity and comparability across studies, to say nothing of saving the effort of developing new measures. On the other hand, replicating results using different measures of the same theoretical concepts can contribute a great deal to confidence in the theory. We presume that informed researchers continue to develop new measures mainly because the old measures have failed to answer their questions and because they believe that some important aspect of the process remains unassessed.

Some beginning researchers may attempt to locate the best or most comprehensive measure of the treatment process. We think that this is impossible. Instead, we suggest, the choice of measure depends on the specific hypothesis, question, or topic being investigated. The examples in this chapter illustrate how investigators choose or design their measures to address each study's specific purposes.

After the measures have been applied, they can be reported directly, as in case studies or intensive analyses of brief segments. More often, measures are aggregated across some stretch of treatment (the summarizing unit), for example, the frequency or percentage of a category in each session, or the average of a rating across a whole treatment. Or the process may be described by using the measures in multivariate or sequential analyses (e.g., Czogalik & Russell, 1995; Luborsky, 1995; Russell, 1995; Russell & Trull, 1986; Stiles & Shapiro, 1995; Stinson, Milbrath, & Horowitz, 1995). The examples in this chapter illustrate some of the alternatives.

## DO THERAPIES DIFFER?

One way to resolve the equivalence paradox is to challenge the assumption that the treatments are different. At one time, some influential authors believed that, despite theoretical differences, the behavior of therapists was much the same across many therapies (e.g., London, 1964). However, process researchers have repeatedly identified systematic differences in therapists' techniques across different orientations (e.g., Brunink & Schroeder, 1979; DeRubeis, Hollon, Evans, & Bemis, 1982; Elliott et al., 1987; Hill, O'Grady, & Elkin, 1992; Hill, Thames, & Rardin, 1979; Startup & Shapiro, 1993; Stiles, 1979; Stiles, Shapiro, & Firth-Cozens, 1988; Strupp, 1955, 1957). The empirically demonstrated process differences have generally been consistent with the theoretical differences between treatments.

To assess differences in treatment processes, investigators have applied process measures to contrasting treatments and compared the results. To illustrate this straightforward logic, we have summarized a few studies that have used different sorts of measures, including verbal category systems, adherence rating scales, and intensive qualitative comparison.

### Treatment Differences in Participants' Verbal Behavior

Verbal response modes (speech act categories, e.g., reflections, interpretations, questions, self-disclosures) have probably been the most widely studied therapist process variable (Elliott et al., 1987; Hill, 1986; Stiles, 1986). In terms of the meta-classification principles, verbal response modes are speech act categories; the scoring units are sentences or clauses; and they are usually coded (into nominal categories) from tapes or transcripts from a trained coder's (i.e., an external observer's) perspective. Most response mode systems are applicable across theoretical orientations and treatment modalities.

#### *Illustrative Study 1*

Stiles et al. (1988) investigated both clients' and therapists' verbal response mode profiles in two different treatment conditions, interpersonal-psychodynamic therapy and cognitive-behavioral therapy. Clients ( $N = 39$ ) were randomly assigned to receive eight weekly sessions of one treatment condition, followed by eight additional weekly sessions in the other treatment condition in a crossover research design that permitted within-subjects comparisons of verbal processes across the two treatments.

Therapists' and clients' verbal behavior was measured using a verbal response modes taxonomy in which each utterance (simple sentence or independent clause) was coded twice, once for its grammatical form (literal meaning) and once for its communicative intent (pragmatic meaning). Utterances were coded into one of eight mutually exclusive categories: disclosure, edification, question, acknowledgment, advisement, confirmation, interpretation, and reflection (Stiles, 1992). To illustrate, "Could you tell me how you felt?" is coded question in form but advisement in intent (i.e., the utterance was apparently meant as a directive, "Tell me how you felt," rather than a question about the client's ability). For each client, four complete sessions of each treatment condition (eight sessions in all) were coded by five coders, who worked directly from audiotape recordings. Coders received 40 to 60 hours of training, and throughout the coding period, they met weekly with investigators to receive feedback and prevent coder drift (the tendency of a coder's standards to change over time). Eighty-four sessions, or approximately one-fourth of all sessions, were coded independently by two coders. This was done to assess intercoder reliabilities of the category percentages. The reliabilities were measured by the intraclass correlation coefficient (Shrout & Fleiss, 1979).

Consistent with cognitive-behavioral therapy's active, directive, and educational stance, therapists averaged 50% more utterances per session in cognitive-behavioral than in interpersonal-psychodynamic therapy, while using higher percentages of questions, edifications (informational statements), and general advisements (directives concerning behavior outside the therapy session). Focusing on client experience and interpersonal relations, the therapists used higher percentages of interpretations

and reflections in interpersonal-psychodynamic therapy. Clients used more acknowledgments (e.g., "mm-hm," "yeah") in prescriptive therapy, and more disclosure in interpersonal-psychodynamic therapy, although these client differences between treatments were much smaller than the therapists' differences. The most common client mode in both treatments was disclosure, consistent with the client role of revealing thoughts and feelings.

### *Illustrative Study 2*

As part of a larger components analysis of cognitive-behavioral self-control therapy with children, Braswell, Kendall, Braith, Carey, and Vye (1985) examined client and therapist verbal behaviors in three treatment conditions. Children ( $N = 27$ ) in grades 3 through 6 were referred by their teachers for having displayed non-self-controlled behavior. Nine students each were assigned to the three treatments, which consisted of 12 45–55 minute sessions over seven weeks. The cognitive-behavioral treatment consisted of verbal self-instruction via modeling with a response-cost contingency for promoting appropriate task performance and self-instruction. The behavioral treatment used modeling and behavioral contingencies, and the attention-control treatment used the same tasks and materials but did not receive either self-instructional training or behavioral contingencies. Each session was audiotaped, and sessions 1, 2, 3, 10, 11, and 12 were analyzed.

The coding system, developed for the study, included 15 categories: five for children's behaviors (self-disclosure, suggested change in task or procedure, evaluative statement about own performance, statements unrelated to task at hand, and duration of the verbal behavior unrelated to the task), nine for therapist behaviors (self-disclosure, emphasizing feelings, correcting the child regarding use of self-instructions, rate of speed or attentiveness, verbal positive reinforcement, asking for feedback on task difficulty or level of enjoyment, frequency of statements unrelated to the task at hand, and duration of verbal behavior unrelated to the task), and one joint therapist-child behavior (duration of task-related activity). The coding unit was the statement or utterance.

Fourteen undergraduate coders were split into three coding groups and were trained for a total of 24 hours in biweekly sessions. When coders achieved 85% agreement on codes, they began coding the study data. Mean reliabilities for each code ranged from 89% to 100% agreement.

Comparisons among treatments indicated that in-session verbal behavior was consistent with treatment style. Cognitive-behavioral and behavior treatment groups showed higher levels of child positive statements about performance, therapist corrections of child performance, and on-task duration. The three groups did not differ in the frequency of therapist encouragement or confirmation. Additional analyses linked the children's active and positive involvement with better outcomes.

### **Treatment Differences in Adherence Ratings**

To ensure treatment integrity in clinical trials comparing different treatments, researchers have tried to standardize the treatments using detailed treatment manuals (DeRubeis et al., 1982; Luborsky, Woody, McLellan, O'Brien, & Rosenzweig, 1982). This step has led researchers to assess therapists' adherence to therapeutic protocols.

