Cognitive Therapy and Research, Vol. 23, No. 3, 1999, pp. 271-283

Inter-Rater Reliability of Cognitive–Behavioral Case Formulations of Depression: A Replication

Jacqueline B. Persons[1,3] **and Andrew Bertagnolli**[2]

We developed a model of cognitive–behavioral case formulation and tested several hypotheses about therapists' ability to use it to obtain cognitive–behavioral formulations of cases of depressed patients. We tested whether clinicians, using measures we developed, could correctly identify patients' overt problems and agree on assessments of patients' underlying schemas. Clinicians offered cognitive–behavioral formulations for three cases after listening to audiotapes of initial interviews with depressed women conducted by the first author in her private practice. Therapists identified 67% of patients' overt problems. When schema ratings were averaged over five judges, inter-rater reliability was good (inter-rater reliability coefficients averaged 0.72); single judges showed poor inter-rater agreement on schema ratings (inter-rater reliability coefficients averaged 0.37). Providing therapists with a specific context in which to make ratings did not improve schema agreement. Ph.D.-trained therapists were more accurate than non-Ph.D.-trained therapists in identifying patients' problems. Most findings replicated those obtained in an earlier study.

KEY WORDS: case formulation; inter-rater reliability; schemas.

One goal of cognitive–behavior therapy (CBT) is to solve overt problems by changing cognitions and behaviors. Change in underlying cognitions, or *schemas,* is also considered quite important, both in the process of treating overt problems and to prevent relapse. Therefore, reliable methods for assessing patients' overt problems and underlying schemas are needed. The importance of case formulation to the practice of CBT is reflected in the fact that the newest measure of cognitive therapy adherence includes items intended to assess the therapist's use of an individualized formulation (Liese, 1995).

Persons (1989, 1993a; Persons & Tompkins, 1997) developed a framework for conceptualizing cases from a cognitive–behavioral point of view. Cognitive–Behavioral Case Formulation emphasizes the importance of identifying the patient's

[1] University of California, San Francisco, and San Francisco Bay Area Center for Cognitive Therapy, Oakland, California.

[2] California School of Professional Psychology, Alameda/Berkeley, California.

[3] Please direct correspondence to Dr. Persons at the San Francisco Bay Area Center for Cognitive Therapy, 5435 College Avenue, Oakland, California 94618.

overt problems and specifying the underlying schemas, or core beliefs, that, when activated by life events, are postulated to cause the overt problems (cf. Beck, Rush, Shaw, & Emery, 1979).

The Cognitive–Behavioral (CB) Case Formulation model asks therapists to make a list of the patient's overt problems; these are concrete difficulties, such as depressive symptoms, fear of freeway driving, social anxiety, binge eating, legal problems, financial difficulties, and interpersonal conflicts. Using the CB Case Formulation model, therapists make a comprehensive problem list, identifying both the problems the patient asks for help with as well as others that the patient may not mention. The need for a comprehensive problem list is based on the notion that if the therapist knows about not only the patient's stated presenting problem, but also of other problems that the patient may have but may not spontaneously report (see also Nezu & Nezu, 1993; Surber, 1994; Turkat & Maisto, 1985). For example, depressed patients often abuse substances; if the therapist treating a depressed patient is not aware of the patient's substance abuse, this problem can undermine the depression treatment. In the present study, we test the hypothesis that therapists, following brief training that emphasizes the importance of a comprehensive problem list and provides some guidelines for making a problem list, can make a comprehensive problem list for a patient.

The cognitive–behavior therapist also identifies schemas, or core beliefs, that the therapist hypothesizes underpin and cause the overt problems when activated by life events or situations. In the CB Case Formulation, therapists identify the patient's views of self, others, and the world. In the present study, we test the hypothesis that therapists, following some brief training, can agree on ratings of schemas for a particular patient. We assess whether therapists can agree on schemas rather than whether their schema ratings are accurate because no criterion measure of a person's schemas is available.

Few investigators have studied cognitive–behavioral case conceptualization. Beckham et al. (1984) showed that therapists were 76% accurate in identifying, for a particular patient (four patients were studied), the underlying schemas chosen by another team of clinicians as characteristic of that patient. Muran and colleagues (Muran & Segal, 1992; Muran, Segal, & Samstag, 1994) developed an idiographic assessment of patients' self-schemas based on the cognitive model; this model focuses only on the patient's views of self. In an earlier study (Persons, Mooney, & Padesky, 1991), we found that clinicians usually identified 65% or more of patients' overt problems, and when groups of five judges were averaged, reliability coefficients reflecting agreement on schema ratings averaged .76. Inter-rater reliability of schema identification was poor for single judges (reliability coefficients averaged .46).

The present study was conducted with the hope of increasing the reliability and validity ratings obtained in our earlier study. To improve therapists' ability to identify patients' overt problems, we taught them to consider a specific list of problem domains when making a problem list, using a list based on work by Nezu and Nezu (1993). The problem domains were: psychiatric symptoms and problems (e.g., depressive symptoms, panic attacks); interpersonal problems; work difficulties; financial difficulties; health problems; housing problems; and recreational difficulties.

To improve schema ratings, we added anchor points to the rating scale and provided more examples in our teaching. We also offered clinicians some specific contexts to consider when they made their schema ratings; that is, we asked clinicians to make schema ratings for a patient who had a public speaking anxiety by considering what the patient's views of self, others, and the world might be in that particular situation. We predicted that clinicians whould be more likely to agree on schema ratings when ratings were made in a specific context than when no context was provided. This prediction was based on the notion that the context, which was chosen because it was problematic for the patient, might provide some initial hypotheses to clinicians about the types of schemas that are commonly activated in that situation (e.g., a public speaking situation commonly activates "self" schemas about inadequacy and "other" schemas about criticism).

What determines a therapist's accuracy in identifying problems and agreement with other clinicians on schema ratings? The answer to this question has implications for training and selection of therapists. We expected that clinicians with Ph.D.-level training might have more specialized training in a wide range of related tasks and skills, and thus might perform better. We expected that clinicians with previous training in case formulation of any type might perform better on this task. We also expected that those with specialized cognitive, behavioral, or CB methods or who use CBT methods more might find the tasks more familiar and easier and might, therefore, perform better. We expected that clinicians with more experience might have had more practice with these or similar tasks and might, therefore, perform better. We collected demographic and training information from the therapists to test these hypotheses.

In summary, we have developed a model of CB case formulation that calls for the therapist to identify the patient's overt problems and schemas likely to underly those problems. In this study, we tested the hypotheses that, using this model and the insstruments developed here, and following a brief (2 ½ hours) training, clinicians can accurately identify patients' overt problems and can agree with one another on ratings of patients' schemas about themselves, others, and the world. We tested the hypothesis that therapists would agree more on schemas when schema ratings while considering the patient in a specific context than when no context was provided. We also tested the hypotheses that Ph.D.-level training, training in case formulation, training in CBT, and clinical experience would improve clinicians' performance on these tasks.

METHOD

Subjects

Clinician subjects were 47 mental health professionals who participated in a day-long training/research workshop in CB case formulation conducted by the first author. Nine subjects were clinicians who attended the workshop when it was given at the annual convention of the Association for Advancement of Behavior Therapy in Atlanta, Georgia, in November, 1993. Thirty-eight subjects were clinicians who attended the workshop when it was given at the V.A. Medical Center in Palo

Alto–Menlo Park in July 1994. Forty-seven mental health professionals attended the Palo Alto session; data from four subjects were discarded because they had no clinical experience (they were researchers or administrators) and data from five subjects were discarded because they were incomplete; therefore, thirty-eight clinicians provided complete data at the Palo Alto site. Because all clinicians received the same training and provided the same measures, data from the Atlanta and Palo Alto samples were combined. Demographic and training characteristics of the clinicians are presented in Table I.

Patient subjects were two depressed and anxious women (''Megan'' and ''Lisa'') treated by the first author in her private practice. A third case served as a practice case (this was the first case studied in Persons et al., 1995; ''Megan'' and ''Lisa'' have not been studied before). All patients gave written permission allowing their therapy sessions to be studied. The practice case was a 23-year-old student who met Axis I criteria for Major Depression and Generalized Anxiety Disorder. Megan was a 32-year-old inventory manager at a large department store who was living with her boyfriend. She met criteria for Major Depression, Dysthymia, and Personality Disorder NOS (avoidant and passive–aggressive features). Lisa was a 56-year-old housewife who was living with her husband. She met criteria for Major Depression, Dysthymia, Social Phobia, Undifferentiated Somatoform Disorder (multiple physical complaints not fully explained by a known medical condition), Dependent Personality Disorder, and Avoidant Personality Disorder. Patients are described more fully in the Results section titled ''Obtaining a Criterion Problem List.''

Measures

Problem List

Clinicians were asked to list patients' overt problems and to provide a few words of detail about each problem. Clinicians were given space to list a maximum of eight problems for each case, in a free-response format.

Table I. Demographic and Training Characteristics of Clinicians ($N = 47$)

Characteristic	Mean or % (*SD*)
Percent female	66.7[a]
Highest degree	
Percent Ph.D.	44.7
Percent M.A. or M.S.W.	44.7
Percent B.A.	10.6
Percent students	12.8
Unlicensed	19.0[a]
Percent with previous training in case formulation	63.0
Hours training in CBT case formulation	173.9 (589.5)[a]
Hours training in cognitive therapy (CT), behavior therapy (BT), or CBT	1290.9 (3206.2)[a]
Hours/week doing CBT	6.0 (8.1)[b]
Years of clinical experience	10.0 (7.8)

[a] $n = 42$.
[b] $n = 45$.

Schemas

A multiple-choice questionnaire assessed clinicians' judgments about each patient's views of self, others, and the world. The questionnaire listed 15 adjectives describing the client's view of self, others, and the world. Clinicians were asked, "Please rate the strength of (patient's pseudonym)'s belief in each item using this scale from 0 to 10," where the 0 point on the scale was labeled "no belief" and 10 was labeled "very strong belief."

Adjectives describing self, others, and world were as follows. Self: defective; wonderful; passive; special; weak, fragile; strong; inadequate; entitled; unimportant; no good; responsible for others; bad; incompetent; unable to cope on my own; undeserving, unworthy. Others: unsupportive; strong; weak; supportive, helpful; dominating, controlling; important; critical; abusive; abandoning; treating me unfairly; unavailable; stupid; passive; unconcerned about me; self-centered. The world: bad; predictable; cruel; benevolent; dangerous; malevolent; overwhelming; negative; unfair; unpredictable; empty, purposeless; potentially catastrophic; fulfilling; unrewarding; challenging. These items were selected from a larger set of items used by the first author in her formulations in a set of approximately 50 cases of depressed outpatients treated in her practice and from items used in a previous study (Persons et al., 1995).

Clinicians provided three sets of schema ratings for Megan and Lisa. Clinicians rated these patients' schemas without any context instructions and in two specific contexts. (Clinicians were not given any context instructions for the practice case.) The two specific contexts for Megan were: "When Megan is at work, functioning as a manager" and "When Megan is interacting with her boyfriend." The two contexts for Lisa were: "When Lisa is in a public-speaking situation" and "When Lisa is interacting with her husband."

Demographics and Training

A brief questionnaire asked clinicians to provide information about demographic characteristics, training, and clinical experience.

Procedure

In the morning of the workshop day, the first author presented didactic material on CB Case Formulation. Next, to practice the formulation process, clinicians listened to an audiotape of the first 12 minutes of an initial session conducted with a practice case by the first author and completed the case formulation measures described previously. Then the first author provided some feedback about the case and the formulation.

In the afternoon, clinicians listened to audiotapes of two initial sessions (Megan and Lisa) of CBT conducted by the first author and completed the case formulation measures described previously. Audiotapes were edited to delete identifying information, segments in which the interviewer summarized the problem list or formulation, and redundancies; each audiotaped segment was about 35 minutes long. When listening to the audiotape, raters also had a typed transcript of the audiotape.

After receiving some feedback about the cases, participants completed demographic and workshop evaluation questionnaires.

RESULTS

We tested four hypotheses: (1) clinicians can accurately identify patients' overt problems; (2) clinicians can agree with one another on ratings of schemas underpinning a patient's overt problems; (3) clinicians agree more on schema ratings when ratings are made in a specific context than when no context is provided; (4) clinicians with Ph.D.-level training, training in case formulation, training in CBT, or more clinical experience perform better on these tasks than those without Ph.D.-level training, with less training in case formulation, less CBT training, and with less experience.

Identification of Overt Problems

To test the hypothesis that clinicians can accurately identify patients' overt problems, we calculated the proportion of clinicians who recognized the problems listed on a criterion problem list for each case.

Obtaining Criterion Problem Lists

The criterion problem list for the practice case was developed in a previous study (Persons et al., 1995) and was based on judgments of two experts (J. Persons and K. Mooney). Criterion problem lists for the cases of ''Megan'' and ''Lisa'' were developed by three clinicians (the authors of the present study and a graduate student). Information used to develop the criterion problem lists included the first author's extensive knowledge of the cases based on her treatment of both patients and pilot work in which six therapists in training and nine practicing clinicians provided problem lists for both cases.

Criterion Problem Lists

The criterion problem list for the practice case had three items (family problems, guilt, and social isolation). This list was shorter than the list for the other cases because it was based on only the first 12 minutes of the initial interview rather than on the entire interview, as was done for the cases of Megan and Lisa.

The criterion problem list for Megan had eight items: work difficulties; difficulties in relationship with boyfriend; depression/anxiety; ''escaping,'' procrastination; not pursuing creative interests; difficulties in relationships with friends; smoking; avoiding driving. The criterion problem list for Lisa had six items: fatigue, frequent illnesses; depression/resentment; generalized anxiety; social anxiety; marital difficulties; interpersonal difficulties (unassertiveness, conflict).

Scoring Clinicians' Problem Lists

For each problem on each criterion list, clinicians received a score of 1 if their problem list included that problem and 0 if it did not. Scoring was generous; clinicians received a score of 1 if the problem in question occurred anywhere (even as a subproblem of another problem or as an aside) on the clinician's problem list. The decision rule used by raters to determine whether the clinician had recognized the problem was: "If I were supervising this clinician with this case, would I feel that the clinician was "getting" the problem?"

Inter-rater reliability of raters' scoring of clinicians' problem lists was high. For the Atlanta sample, the two authors scored the problem lists for Lisa for the first four subjects and then compared ratings and refined their scoring criteria. Then they scored all the remaining subjects' problems lists for all cases; the raters agreed 87% on those ratings. For the Palo Alto sample, the two authors scored the problem lists for all three cases provided by six randomly selected clinician subjects. The two judges agreed on 93% of ratings and therefore the second author scored the problem lists for the remaining clinician subjects.

Clinicians' Identification of Criterion Problems

The practice case had three problems, Megan had eight, and Lisa six, for a total of 17 problems across all three cases. On average, clinicians rated 16.23 ($SD = 2.10$) problems (a few clinicians did not rate one of the cases). On average, of the 16.23 problems they rated, clinicians correctly identified 10.94 ($SD = 2.59$) problems. Of the problems rated, the average percentage correctly identified was 67.46% ($SD = 13\%$). Thus, clinicians correctly identified about two-thirds of the problems they rated.

Inter-Rater Reliability of Schema Ratings

To test the hypothesis that clinicians can agree on schema ratings, we assessed inter-rater reliability of schema ratings by calculating intraclass correlation coefficients (ICC; Shrout & Fleiss, 1979) separately for each case for each category of schema (views of self, other, and world) for each case and each context for each case. The ICC is essentially a ratio of the proportion of variance in ratings due to "targets" divided by the sum of the proportion of variance due to "targets" plus the porportion of variance due to "judges." If the proportion of variance due to judges is low and the proportion of variance due to targets is high, then the ICC approaches 1 and inter-judge reliability is high.

In our analyses, the "judges" were the clinicians who provided ratings, and the "targets" were not individuals, as in the usual ICC computation; instead, targets were the individual items in each category (views of self, other, world). For example, the "targets" in the ICC analysis for "self" are the 15 adjectives that describe the self. Our ICC, thus, is essentially a ratio of the proportion of variance in ratings due to "items" divided by the sum of the proportion of variance due to "items" plus the proportion of variance due to "judges." If the variance due to judges is low and the variance due to items is high, then the inter-rater reliability is high. If the variance due to judges is high, then the inter-rater reliability is low.

The repeated-measures analysis of variance that underlies the ICC computations assumes independence from target to target. However, with items replacing targets, it is likely that there is some correlation among items. However, the independence assumption is needed only for statistical inference, and we are using the ICC here as a descriptive statistic. The method used here is also the method adopted by the Mount Zion researchers (Curtis et al., 1988; Rosenberg et al., 1986) and in our own previous work (Persons et al., 1995).

Table II presents ICCs for each case and type of rating for a single, random judge and for a mean of a random sample of five judges. The ICC for a single judge is the estimated ratio of variance due to targets to the sum of variance due to targets and judges (even though it is for a single judge). When more than one judge is used, the variance due to judges goes down, so reliability goes up. To say this another way: As the number of judges upon which a rating is based increases (from one to five), the reliability of the rating increases (Horowitz et al., 1989). We chose the figure five because clinical meetings held to discuss and formulate a case might involve a group of that size.

As Table II shows, inter-rater reliability coefficients were good for five judges (ranging from 0.44 to 0.91 and averaging 0.72) and poor for single judges (ranging from 0.13 to 0.66 and averaging 0.37). These figures were very similar to those obtained in a previous study of the practice case (inter-rater reliability coefficients averaged 0.46 for single judges and 0.80 when averaged over five judges).

Effects of Context on Schema Agreement

To test the hypothesis that inter-rater agreement would be higher when specific contexts were provided than when they were not, an analysis of variance using *z*-

Table II. Inter-Rater Reliability for Clinicians' ($N = 47$) Judgments of Schemas of Self, Other, and World for Three Cases in General and Specific Contexts

	Single judge	Five judges
Practice Case		
Views of self	0.35	0.73
Views of others	0.55	0.86
Views of world	0.34	0.72
Megan		
Views of self—general	0.50	0.83
Views of self—manager context	0.28	0.66
Views of self—boyfriend context	0.25	0.63
Views of others—general	0.24	0.61
Views of others—manager context	0.17	0.51
Views of others—boyfriend context	0.35	0.73
Views of world—general	0.31	0.70
Views of world—manager context	0.20	0.56
Views of world—boyfriend context	0.33	0.71
Lisa		
Views of self—general	0.55	0.86
Views of self—public speaking context	0.66	0.91
Views of self—husband context	0.38	0.75
Views of others—general	0.38	0.75
Views of others—public speaking context	0.39	0.76
Views of others—husband context	0.62	0.89
Views of world—general	0.37	0.75
Views of world—public speaking context	0.40	0.77
Views of world—husband context	0.13	0.44

Note: Intraclass correlation coefficients (Shrout & Fleiss, 1979) are presented.

transformed ICC values was computed. Independent variables were CASE (practice, Megan, Lisa), VIEW (self, other, world), and CONTEXT (specific context, no context). An interaction variable for VIEW × CONTEXT was also entered in the model. The overall *R*-square of the model was 0.59; none of the independent variables or the interaction effect were statistically significant at the $p < .05$ level. Thus, contrary to prediction, raters did not agree more often on schema ratings when specific context ratings of schemas were made than when no context was provided.

Effects of Training and Experience

Problem Identification

We tested the hypothesis that clinicians with Ph.D.-level training, those with more training in case formulation, those with more training in CBT, or those with more clinical experience identify problems more accurately than clinicians without Ph.D.-level training, with less training in case formulation, with less CBT training, or with less experience. To test this hypothesis, we conducted a multiple regression analysis, in which the dependent variable was the logit-transformed proportion of problems correctly identified and the independent variables were: Ph.D. (coded 0-no or 1-yes), prior training in case formulation (coded 0-no or 1-yes), hours training in CB case formulation, hours of training in CBT, hours of weekly CBT provided, and years of clinical experience.

Results of this analysis for an *N* of 38 subjects show that the overall model is statistically significant (*R*-square is 0.34, $p = 0.034$) and only one independent variable, Ph.D.-level training, was statistically significant ($p = 0.019$). The residuals of this model were normally distributed ($p = 0.46$).

Inter-rater Reliability of Schema Ratings

We tested the hypotheses that therapists with Ph.D.-level training, with more training in case formulation, more training in CBT, or with more years of experience were more likely to agree with one another on schema ratings than those without Ph.D.-level training, with less training in case formulation, less training in CBT, or with fewer years of experience. To do this we began, by calculating, for each judge, for each view (self/other/world), for each context, and for each case, an ‘‘agreement index.’’ The ‘‘agreement index’’ is a correlation (Pearson product–moment correlation) between a particular judge's rating and the average of the other judges' ratings, divided by the average correlation among all judges. Average correlations are computed after transforming using Fisher's *Z* transformation. This method is recommended by Williams (1976) and we used it in an earlier study (Persons et al., 1995). We used the ‘‘agreement index’’ rather than the ICC because the ICC for a single judge provides information about the degree to which a single judge agrees, on average, with any other single judge; however, in order to examine predictors of inter-rater reliability, we needed a figure that would measure the degree to which the ratings of each particular judge agreed with the ratings of all of the other judges in the sample.

A multiple regression was conducted using the ‘‘agreement index’’ as the dependent variable and the same independent variables as in the previous analysis. The overall model is not very impressive (R-square $= 0.152$, $p = 0.49$), and residuals are normally distributed ($p = 0.35$). None of the independent variables are statistically significant at the $p < .05$ level. Thus, none of the demographic or training variables predicted clinicians' tendency to agree with the other clinicians on schema ratings.

DISCUSSION

Clinician raters identified, on average, about two-thirds of patients' overt problems. This figure is at first blush a bit disappointing. However, it proves to be quite a bit superior to the figures obtained by other investigators. Hay et al. (1979) studied problem areas rated by four interviewers, each of whom interviewed the same four clients. The mean rate of agreement between interviewers on the presence of specific problem areas was .55 {rate of agreement = agreements/(agreements + disagreements)}. Wilson and Evans (1983) reported that 38.6% of judges selected the most commonly agreed-upon priority target behavior when they reviewed written descriptions of three cases of child psychopathology; a somewhat higher figure (48.2%) was obtained when the proportion of judges identifying the patient's six problems was calculated.

The problem identification rate we obtained in this study is similar to the rates reported in our earlier study (Persons et al., 1995). Although in this study we taught clinicians to consider a list of problem domains, this proved insufficient to increase the problem identification rate. Clinicians might be more accurate at problem identification if they completed a checklist of problem domains when assessing the patient, or patients themselves might be asked to complete such a checklist. The closest available measure of this sort that we are aware of is the Quality of Life Inventory developed by Frisch (1992). The Quality of Life Inventory is a self-report measure that asks individuals to rate their satisfaction in 16 life domains. A limitation of the Quality of Life Inventory is that it measures satisfaction, not functioning.

The importance of comprehensive problem identification and assessment is supported by the work of Linehan (1993); her manual for treating parasuicidal women with borderline personality disorder stresses identification of the full range of these patients' overt problems. Miranda (1995) also reported that assessment and treatment of the multiple problems of disadvantaged depressed medical patients produced better outcome than treatment focused solely on depressive symptoms. Thus, measures of presenting problems are urgently needed.

Inter-rater agreement of clinicians' ratings of patients' schemas was good when ratings were averaged over five judges (mean inter-rater reliability coefficient of 0.72), poor when single judges were considered (averaging 0.37). Certainly it is well known that averaging ratings over multiple judges produces higher agreement than when single judges are examined (cf. Horowitz et al., 1989). This finding suggests that clinicians can benefit from consulting with one another when formulating schema hypotheses about their patients. Consultation with the patient is also useful to enhance reliability (and collaboration).

Judges did not agree more often when rating schemas in a specific context than when no context was provided. Why not? And we were not able to improve inter-rater reliability of schema ratings over our earlier study (Persons et al., 1995). How can this be done? We address these two questions together.

To improve inter-rater reliability of schema ratings, we propose that teachers must list, very explicitly, the typical schemas of patients who have particular presenting problems that occur in particular situations. If this were done, clinicians presented with those presenting problems and situations could agree more often on schema ratings. We speculate that the relatively good inter-rater reliabilities obtained for the Plan Formulation method (Curtis, Silberschatz, Sampson, & Weiss, 1994; Rosenberg et al., 1986) are due at least in part to the fact that the theory underlying the method clearly states how to conceptualize the case (the theory states that patients' problems arise from survival or separation guilt relating to parental figures).

One training variable, earning a Ph.D., predicted clinicians' ability to identify presenting problems. We did not obtain this result in our earlier study; therefore, this finding deserves replication before it can be accepted without reservation. The link between Ph.D.-level training and identification of presenting problems is not straightforward, and the variable Ph.D.-level training most likely serves as a proxy for a number of other factors, possibly including training in diagnostic and psychological assessment of all types.

The present study has several limitations. Although a strength of the study is that it examines data collected in a "real world" clinical setting, the study does not completely reflect some of the processes that "real world" clinicians use to formulate cases. Raters had access to transcripts in addition to the audiotape material; therefore, if they paged backward or forward in the transcript, they processed material differently from the way it is done in a therapy session, when material must be processed in the sequence in which it is received from the patient. Audiotape material does not provide therapists with the visual cues that are useful in assessing patients' problems and schemas, particularly interpersonal ones. In addition, as they formulated the case, therapists were required to follow the interview sequence pursued by the interviewer rather than asking the questions that would have allowed them to develop and test their own clinical hypotheses. The three patients studied were all female and were selected because good audiotapes were available, the patients gave permission to be studied, and the cases seemed relatively straightforward. Clinicians were a convenience sample. As a result, findings of this study do not necessarily generalize to other patients and clinicians.

Overall, clinicians were moderately good at identifying presenting problems and proposing schema hypotheses. An important next step in this line of work is the demonstration that an accurate and reliable individualized formulation contributes to treatment outcome. Some early studies of this question have been disappointing. A study by Emmelkamp, Bouman, and Blaaw (1994) found no outcome superiority for patients who were treated via an individualized formulation-driven treatment as compared to a standardized treatment, and a study by Schulte, Kunzel, Pepping, and Schulte-Bahrenberg (1992) found that standardized treatment was superior to individualized treatment. Certainly the Emmelkamp et al. (1994) result might be accounted for by the low power of the study and both results may be accounted for in part by the fact that patients in these two studies had relatively homogeneous problems. Perhaps an individualized case formulation is particularly important in the treatment of patients with multiple problems. Nevertheless, this has not been shown as yet, and thus findings to date do not provide strong support for importance to outcome of an individualized formulation. More encouragement can be obtained from the findings that depressed patients whose underlying schemas were effectively treated relapsed less often than patients who ended acute treatment for depression with high levels of dysfunctional schemas (Blackburn, Eunson, & Bishop, 1986; Evans et al., 1992; Simons, Murphy, Levine, & Wetzel, 1986). These findings remind us that attention to underlying core schemas may contribute more to relapse prevention than to the outcome of acute treatment.

ACKNOWLEDGMENTS

We thank the patient subjects for giving permission to study their cases and the clinicians for their time. We thank the Association for Advancement of Behavior Therapy, and particularly Mary Jane Eimer, for allowing data to be collected at the convention in 1993, and the Palo Alto Veterans Administration, particularly Antonette Zeiss and Jacqueline Becker, for allowing data to be collected there. We thank Michelle Hatzis and Joan Davidson for participating in the research seminar that guided this work, Bert Epstein for assisting with data collection in Palo Alto, and Alan Bostrom for statistical assistance. This work was supported by grant MH50367 from the National Institute of Mental Health.

REFERENCES

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York: Guilford.

Blackburn, I. M., Eunson, K. M., & Bishop, S. (1986). A two-year naturalistic follow-up of depressed patients treated with cognitive therapy, pharmacotherapy and a combination of both. *Journal of Affective Disorders, 10,* 67-75.

Curtis, J. T., Silberschatz, G., Sampson, H., Weiss, J., & Rosenberg, S. E. (1988). Developing reliable psychodynamic case formulations: An illustration of the Plan Diagnosis method. *Psychotherapy, 25,* 256-265.

Curtis, J. T., Silberschatz, G., Sampson, H., & Weiss, J. (1994). The Plan Formulation Method. *Psychotherapy Research, 4,* 197-207.

Emmelkamp, P. M. G., Bouman, T. K., & Blaauw, E. (1994). Individualized versus standardized therapy: A comparative evaluation with obsessive-compulsive patients. *Clinical Psychology and Psychotherapy, 1,* 95-100.

Evans, M. D., Hollon, S. D., DeRubeis, R. J., Piasecki, J. M., Grove, W. M., Garvey, M. J., & Tuason, V. B. (1992). Differential relapse following cognitive therapy and pharmacotherapy for depression. *Archives of General Psychiatry, 49,* 802-808.

Frisch, M. B., Cornell, J., Villanueva, M., & Retzlaff, P. J. (1992). Clinical validation of the Quality of Life Inventory: A measure of life satisfaction for use in treatment planning and outcome assessment. *Psychological Assessment, 4,* 92-101.

Hay, W. M., Hay, L. R., Angle, H. V., & Nelson, R. O. (1979). The reliability of problem identification in the behavioral interview. *Behavioral Assessment, 1,* 107-118.

Horowitz, L. M., Rosenberg, S. E., Ureno, G., Kalehzan, B. M., & O'Halloran, P. (1989). Psychodynamic formulation, consensual response method, and interpersonal problems. *Journal of Consulting and Clinical Psychology, 57,* 599-606.

Liese, B. S. (1995). *Cognitive Therapy Adherence Scale (CTAS).* Unpublished manuscript.

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder.* New York: Guilford.

Miranda, J. (1995). Treatment of depression for disadvantaged medical patients. Paper presented at Society for Psychotherapy Research, Vancouver, British Columbia, Canada, June 22-25, 1995.

Muran, J. C., & Segal, Z. V. (1992). The development of an idiographic measure of self-schemas: An illustration of the construction and use of self-scenarios. *Psychotherapy, 29,* 524-535.

Nezu, A. M., & Nezu, C. M. (1993). Identifying and selecting target problems for clinical interventions: A problem-solving model. *Psychological Assessment, 5,* 254-263.

Persons, J. B. (1989). *Cognitive therapy in practice: A case formulation approach.* New York: Norton.

Persons, J. B. (1993a). Case conceptualization in cognitive-behavior therapy. In K. T. Kuehlwein & H. Rosen (Eds.), *Cognitive therapy in action: Evolving innovative practice.* (pp. 33-53). San Francisco: Jossey-Bass.

Persons, J. B. (1993b). Outcome of psychotherapy for unipolar depression. In T. R. Giles (Ed.), *Handbook of effective psychotherapy* (pp. 305-323). New York: Plenum.

Persons, J. B., Mooney, K. A., & Padesky, C. A. (1995). Inter-rater reliability of cognitive-behavioral case formulations. *Cognitive Therapy and Research, 19,* 21-34.

Persons, J. B., & Tompkins, M. A. (1997). Cognitive-behavioral case formulation. In T. D. Eells (Ed.), *Handbook of psychotherapy case formulation.* New York: Guilford.

Rosenberg, S. E., Silberschatz, G., Curtis, J. T., Sampson, H., & Weiss, J. (1986). A method for establishing reliability of statements from psychodynamic case formulations. *American Journal of Psychiatry, 143,* 1454-1456.

Schulte, D., Kunzel, R., Pepping, G., & Schulte-Bahrenberg, T. (1992). Tailor-made versus standardized therapy of phobic patients. *Advances in Behaviour Research and Therapy, 14,* 67-92.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Simons, A. D., Murphy, G. E., Levine, J. L., & Wetzel, R. D. (1986). Cognitive therapy and pharmacotherapy for depression. *Archives of General Psychiatry, 43,* 43-49.

Surber, R. W. (Ed.). (1994). *Clinical case management: A guide to comprehensive treatment of serious mental illness.* Thousand Oaks, CA: Sage Publications.

Turkat, I. D., & Maisto, S. A. (1985). Personality disorders: Application of the experimental method to the formulation and modification of personality disorders. In D. H. Barlow (Ed.), *Clinical handbook of psychological disorders: A step-by-step treatment manual* (pp. 502-570). New York: Guilford.

Williams, G. W. (1976). Comparing the joint agreement of several raters with another rater. *Biometrics, 32,* 619-627.

Wilson, F. E., & Evans, I. A. (1983). The reliability of target-behavior selection in behavioral assessment. *Behavioral Assessment, 5,* 15-32.