Community Mental Health Journal, Vol. 22, No. 4, Winter 1986

Do Therapists Bias Their Ratings of Patient Functioning Under Peer Review?

Gerald J. Stahler, Ph.D.
Herbert Rappaport, Ph.D.

ABSTRACT: The present study was an attempt to examine the rating bias of therapists participating in an evaluation of an experimental quality assurance system at a community mental health center. The test program was intended to identify patients who demonstrated lack of progress or poor level of functioning after two months of treatment, and to employ a clinical assessment process by independent clinicians to evaluate problems in the quality of care.

It was believed that the therapists knowledge that they might have their clinical work assessed would lead to biased ratings of more severe symptomatology in their patients. The results of this study partially supported the hypothesis. Patients in the peer review system were rated as more dysfunctional at admission on Psychological Functioning than patients in the control groups. No differences, however, were found on Basic Life Functioning, Anti-Social Behavior, or Mental Processes. The implications for these results relative to psychotherapy research, quality assurance, and program evaluation are discussed.

Therapist judgments of client functioning have formed a cornerstone for much of the psychotherapy outcome research literature (Luborsky et al. 1971) as well as for such other activities as quality assurance systems and program evaluation of mental health services (Tash et al., 1982; Tash and Stahler, 1984). The act of studying and evaluating psychotherapy or mental health services, however, may influence the results of such research. Indeed, experimental artifacts producing biased results have been extensively studied in many areas of psychology (Rosenthal, 1966; and Rosenthal and Rosnow, 1969). The fact that therapists are highly trained does not preclude the possibility that they, too, can

Dr. Stahler is Assistant Vice Provost for Research in Graduate Studies, Temple University, as well as Director of Research and Evaluation for Temple's National Institute for Adolescent Pregnancy and Family Services. Dr. Rappaport is Associate Professor, Clinical Division of Psychology, Temple University.

Reprint requests should be addressed to Dr. Stahler, Room 406, University Services Bldg., Temple University, Philadelphia, PA 19122.

The authors would like to thank Dr. William Tash, Dr. Richard Woy, and Dr. Frank Mcguirk for their involvement and input in the original quality assurance study.

be affected by observer bias effects. Sullivan, for example, called attention to the frequent problem of the idealization of the patient by the therapist (1965, cited in Mullahy, 1973). Others have also discussed the biased perceptions of the therapist because of counter-transference (Freud, 1949; Berman, 1959; Cutler, 1958; Epstein and Feiner, 1979; Garfield and Affleck, 1960; Reich, 1951; Searles, 1979; Zirkle, 1959).

Traditionally, however, researchers have tended to perceive the therapist as the best judge of therapeutic progress and efficacy since the clinician is the "expert" and the patient's judgment is distorted by transference (Horenstein, Houston, and Holmes, 1973) or by the need to feel that he/she has not "wasted his time" (Berg, 1952). Thus, in one review of 165 psychotherapy outcome studies, the vast majority utilized *only therapist* rating measures (Luborsky, et al., 1971). Yet it is quite likely that there are biases in the therapist's vantage point.

Rosenberg's work (1969) on evaluation apprehension and social desirability in the psychological experiment is of particular importance in this context. His research suggests that when the subject perceives the experimenter as having "power" over him/her in regard to controlling access to a subject's desired goal or activity, the subject's responses will be biased in the cued direction. In a quality assurance situation or other forms of program evaluation, senior clinical staff or administrators are often viewed as controlling access to such goals as validation of competence, approval, and promotion. Thus, scrutiny of a therapist's clinical work in a peer review or program evaluation situation may arouse evaluation apprehension in therapists and may affect the way therapists report patient symptomatology or treatment progress.

In the psychotherapy outcome research literature, there has been some speculation that rating bias may affect the results of psychotherapy evaluations, but this seldom has been studied. Most investigators have reasoned that when psychotherapists and patients submit the results of their work to the scrutiny of a researcher, they bias their assessments in the direction of successful results (Steinhelber, 1970). Others have reported a "reverse bias" effect in which therapist ratings of patient dysfunction are biased toward less improvement. For example, Parloff, Kelman, and Frank (1954) believe that this may occur when patients are better able to admit psychopathology over the course of treatment. Meehl (1959), Dailey (1952), and Michaux et al. (1963) all found evidence suggesting that therapists tended to underestimate improvement in their patients and overestimate pathology.

Consistent with Rosenberg's research, the bias in the "cued direction" may be toward rating patients as more dysfunctional since the training of psychotherapists and their role functions are largely oriented toward identifying psychopathology so as to more effectively intervene. Thus, in terms of receiving esteem from peers, it may be preferable to be over-zealous in identifying symptomatology than to underestimate or "miss" areas of dysfunctional behavior. This may be particularly true when ratings are conducted as an experiment and are not part of the routine procedures of a treatment facility.

The present study was an attempt to examine therapist rating bias within the context of an evaluation of a demonstration quality assurance system being

tested in a federally funded community mental health center. The primary question under investigation was whether a quality assurance program designed to identify problem cases would result in a biased rating of patient level of functioning.

The experimental quality assurance system involved utilizing a peer review of cases based on level of patient dysfunction or lack of progress after two months of treatment. It was believed that because staff involved in the study knew that their clinical work and their "problem cases," in particular, might come under peer review, their ratings of experimental group patient level of functioning would be more severe at intake and termination than staff ratings of the two control groups which were comprised of randomly selected clients who were not part of this quality assurance program but who did receive treatment in the center.

It was also predicted that the Concurrent Control Group (clients in the center not involved in the demonstration quality assurance program but in treatment in the mental health center at the same time as the Experimental Group) would be rated as more poorly functioning than the Pre-Study Control Group (comprised of clients who had completed treatment in the center prior to the start of the quality assurance project). This would suggest system-wide impact attributable to having a study team in the Center.

METHOD

Subjects

Experimental Group (E Group). Subjects in this group consisted of all new admissions and re-admissions enrolled in adult, adolescent, and drug outpatient therapy, residential treatment, and partial care modalities during alternate weeks over a four month period. Children under sixteen years of age, methadone patients, inpatients, non-English speaking individuals, and those unable to complete the admission forms because of the severity or acuteness of their illness were excluded. This resulted in obtaining 256 subjects for this group, with a mean age of about 30, a mean educational attainment of 12.64 years, and an average individual income of $6,336. Most patients were outpatients (89.5%) and Caucasian (94.1%), with only a slightly higher percentage of females (53.3%) relative to males (46.5%). The majority of subjects were new admissions (79.3%), with only 20.7% having received prior treatment at the mental health center.

Concurrent Control Group (C Group). With the exception of children under sixteen, methadone clients, inpatients, and those unable to complete the admission forms, all patients admitted or re-admitted on alternate weeks during the first three months of the study comprised the Concurrent Control Group. This consisted of 283 patients, with a mean age of 31, mostly Caucasian (91.5%), and an average educational attainment of 12.54 years. The mean individual income for this group was $8,930, and there were slightly fewer males (46.3%) relative to females (53.7%). The majority of the group were new admissions (74.2%), with only 25.8% having received treatment at the center previously.

Pre-Study Control Group (P Group). Since it was suspected that the experimental quality assurance program might have system-wide effects which could influence therapists and/or patients in the Concurrent Control Group, a second control group was included. The Pre-Study Control Group consisted of data obtained from 278 randomly selected patient records meeting the same criteria as the two other groups, but who completed treatment before the start of the present study. Prior to the present study, cases were selected for peer review on a random basis. Subjects were selected using all patients admitted during alternate weeks for a four month period beginning eight months prior

to the start of the study. The group had a mean age of 28.94 years, an average educational attainment of 12.67 years, and a mean individual income of $6,984. A slight majority were female (59.9%), and most were Caucasian (89.9%). Seventy percent were new admissions with 29.9 percent having received prior services in the center.

Demographic equivalence among groups was tested with one way ANOVA's and chi-square analyses on ethnicity, gender, marital status, diagnosis, age, educational attainment, and average annual income. Income was found to be significantly higher in the C-Group, which also had the largest number of missing data on the variable and the highest proportion of subjects with unknown income. It may be speculated that those subjects not reporting income may have had lower income levels than those whose incomes were known.

Instruments

The Level of Functioning scale (LOF) is a nine item rating scale assessing the patient's current level of functioning and is routinely completed by the therapist after the clinical intake session (Pre-Measure) and at termination (Post-Measure) at the center. For the present study, therapists also completed this form on patients in the Experimental Group after two months of treatment (Interim Measure) to assist in determining if peer review was required.

The scale was developed by Carter and Newman (1976) and modifed by the Colorado Division of Mental Health, Statistical Analysis and Research Section, and the Colorado Treatment Outcome Task Force (Edward's McGuirk, and Wilson, 1978: Ellis, 1977) for use in all state mental health agencies and state hospitals in Colorado. The instrument is a 50 point rating scale divided into five levels of ten points each, with zero to ten representing good functioning and 40 to 50 classified as severe disruption. There are nine dimensions on the scale which consist of the following: socio-legal functioning, substance use, medical/physical health, mental processes, emotional health, personal behavior, inter-personal relationships, occupation/education/home managements, and meeting basic needs. This scale has shown moderate criterion and construct validity (Irving, 1981; Krowinsky and Fitt, 1978; and Newman, 1980). All clinicians involved in the study were thoroughly trained in the use of the scale since it has been a standard form in the center for over a year. Prior to making any statistical comparisons, this scale was factor analyzed using an orthogonal (Varimax) rotation to reduce its dimensionality. The factor analysis, based on an n = 809, yielded three dimensions—Basic Life Functioning, Psychological Functioning, and Anti-Social Behavior. The item, "Mental Processes," did not load differentially on the other factors and was kept as a separate item for statistical analysis.

Procedure

All clinicians in the mental health center were briefed in a series of meetings by the Center Director and Director of Evaluation about the new procedure for a proposed study concerning quality assurance. The staff were informed that the study was to be jointly conducted with a consulting firm under contract with the National Institute of Mental Health, and that researchers from the firm would oversee all phases of the project. In addition, the methodology of the study was explained, and a brief memo describing the project was distributed to all personnel. No staff were informed, however, about the present study concerning the examination of therapist rating bias.

Alternate weeks over a four month period were designated as E sampling weeks, during which time all patients admitted or re-admitted on these days would be placed in the Experimental Group. Patients admitted or re-admitted during alternate weeks were included in the Concurrent Control (C) Group. On admission, clients in the E Group were given an admission form and told they might be requested to come in at a later date to be interviewed as part of the center's quality assurance program. C Group clients were given only the admission form to complete. After this administrative intake, clients were assigned to a clinician based on availability and scheduled for a clinical intake session. After meeting with clients for one session, clinicians completed the LOF pre-

measure which is a routine procedure in the center. After two months of treatment E Group clients were again re-evaluated on the LOF. Those showing the least progress or those most dysfunctional according to these measures were referred to the peer review committee, which assessed the process of treatment. This committee reviewed the case file and had the option to utilize clinical assessment interviews by independent center clinicians to assist in identifying problems in the quality of care. After the peer review process, recommendations were then provided to the client's therapist. During the study a total of 90 clients were referred to the peer review committee and 54 clients were invited for assessment interviews. The objective of this procedure was to identify and remediate any deficiencies in the quality of treatment which may have contributed to the lack of client progress, to provide feedback to therapists on their "problem" cases, and to implement a more direct evaluative mechanism for assuring quality of care. At termination, each therapist again completed the LOF for clients in both groups. For the purposes of this study only intake and termination LOF data were used.

RESULTS

It was predicted that therapists involved with patients in the E Group would rate them as more dysfunctional in comparison to those in the C and P Groups. In addition, it was anticipated that C Group Patients would be rated as more dysfunctional than P Group patients indicating a system-wide impact of the demonstration quality assurance program. Two series of analyses were performed on the LOF data. First, planned orthogonal contrasts (Winer, 1971) were utilized to compare the three subject groups on LOF scores at admission. The second series of tests involved using analyses of covariance (ANCOVA) on termination scores corrected for the admission ratings. With variables on which significant differences were found among groups in the ANCOVA, contrasts were employed as post hoc tests with Dunn-Bonferoni corrections. To reduce within cell variance, the E Group was divided into seven groups for purposes of analysis according to the processes of the quality assurance system, but are aggregated here for clarity.

Admission Scores

Planned orthogonal contrasts were used to test for differences between E Group and C and P Groups on admission LOF scores for the three LOF factor scales and Mental Processes item. Table 1 presents the means and standard deviations for admission LOF scores for all groups. Consistent with expectations, E Group ratings of Psychological Functioning were found to be higher (i.e., more dysfunctional) than those of the C and P Groups ($F = 6.60$, $df = 1, 804$, $p < .01$).

On the other hand, no differences were found between the E Group and the C and P Groups on therapist ratings of patient level of functioning on Basic Life Functioning ($F = 1.35$, $df = 1, 803$, $p = .60$), Anti-Social Behavior ($F = 1.35$, $df = 1, 803$, $p = .24$), and Mental Processes ($F = .77$, $df = 1, 805$, $p = .37$) at admission.

In addition, planned orthogonal contrasts were also employed to test whether the C and P Groups differed on LOF ratings to determine the extent of a system-

Table 1
Means and Standard Deviations for Admission LOF Scores For All Subject Groups

Group	*n*	*Basic Life Functioning*	*Psychological Functioning*	*Anti-Social Behavior*	*Mental Processes*
Experimental Subgroups	252	62.12 (28.04)	92.57 (22.74)	41.01 (21.04)	21.54 (10.78)
Concurrent Control	281	60.76 (26.27)	87.42 (22.51)	38.51 (20.71)	21.46 (9.98)
Pre-Study Control	278	59.74 (28.72)	87.42 (24.72)	37.93 (21.68)	22.73 (11.67)
Total	811	60.83	89.02	39.10	21.92

wide effect. No differences were found on Basic Life Functioning ($F = 0$, $df = 1, 804$, $p = 1.0$), Anti-Social Behavior ($F = .10$, $df = 1, 803$, $p = .74$), and Mental Processes ($F = 1.91$, $df = 1, 805$, $p = .16$), thereby indicating no support for the predicted system-wide impact of the quality assurance program.

Termination Scores

Analysis of covariance (ANCOVA) was used to test for differences among groups at termination with values corrected for admission scores. Table 2 presents the means and standard deviations for the LOF at termination. Contrary to expectations, no differences among groups were found on Mental Processes ($F = .56$, $df = 8, 463$, $p = .81$). Significant differences were found, however, among groups on Basic Life Functioning ($F = 2.39$, $df = 8, 463$, $p = .01$), Psychological Functioning ($F = 2.49$, $df = 8, 463$, $p < .01$), and Anti-Social Behavior ($F = 2.21$, $df = 8, 463$, $p = .02$). Contrasts with the Dunn-Bonferoni adjustment were employed to make post hoc comparisons of interest relative to hypotheses: Contract 1 = E vs. C + P; Contrast 2 = C vs. P. This procedure revealed no differences on Basic Life Functioning ($F = .391$, $df = 1, 463$, $p = .53$; $F = .01$, $df = 1, 463$, $p = .89$), $F = .44$, $df = 1, 463$, $p = .50$), and Anti-Social Behavior ($F = .35$, $df = 1, 463$, $p = .55$; $F = .11$, $df = 1, 463$, $p = .74$). The only post hoc tests that revealed significant differences among groups concerned non-hypothesis related comparisons.

Therapist Debriefing Questionnaire

At the conclusion of this study, therapists who had patients as subjects in the study were administered a questionnaire concerning a number of aspects regarding the quality assurance program. One item that was relevant to the present study was the following: "How did the knowledge that some of your more dysfunctional cases might come under peer review and clinical assessment to evaluate the quality of treatment affect you? Did you do anything differently because of this possibility?" Of the 51 therapists given the questionnaire, twenty returned them. These twenty clinicians were responsible for treating 91 of the

Table 2
Means and Standard Deviations for Termination LOF Scores

Group	*n*	*Basic Life Functioning*	*Psychological Functioning*	*Anti-Social Behavior*	*Mental Processes*
Experimental	129	59.93 (25.69)	82.26 (24.78)	40.81 (20.36)	19.41 (9.30)
Concurrent Control	143	56.08 (26.61)	79.79 (26.55)	36.57 (21.19)	19.40 (10.20)
Pre-Study Control	200	55.13 (27.89)	78.70 (28.08)	35.17 (20.88)	20.57 (10.87)
Total	473	56.73	80.00	37.14	19.90

173 E Group patients who had three or more sessions as well as 32 of the 38 patients who were clinically assessed. These clinicians were relatively experienced, having worked for the center for an average of 4.5 years, and having worked professionally as therapists for an average of nine years. Overall, most clinicians reported that the peer review system did not affect them. One clinician mentioned that it made him anxious at first: "I felt anxious early on about this, but now feel little or no anxiety. Peer Review . . . is clearly of a non-punitive nature." Several therapists, however, seemed somewhat defensive in their remarks about a peer review *not* affecting them:

—"I am never threatened by the prospect of my peer reviewing my work as I see this as a learning experience, *not competition*!"

"It was impossible to do any better than I was already doing, as I always do the best I can. It felt critical, but I think it should not be."

"I always do the best I can."

Finally, several welcomed the opportunity of potentially obtaining feedback on their work:

"I was always eager to compare their assessment with my own."

"It felt helpful to have others' opinions."

"It is helpful to have my more dysfunctional cases reviewed."

"I welcomed the opportunity to have the case reviewed, but felt in some ways it placed the client in a bind."

Thus, if therapist self-report is taken at face value, there is little overt evidence to support the prediction that the possibility of peer review changed their ratings of patient symptomatology, although in some cases, therapists had reported a modest level of evaluation apprehension regarding peer review. Nevertheless, considering that the study was conducted in one CMHC, the return rate of therapist questionnaires (40%) may be indicative of the defensiveness we were trying to measure. Although the study was conducted at only one

site, the center is fairly typical of CMHCs nationally, which is one of the reasons the site was selected. However, additional invesigation in other centers will be needed to further assess the validity of these results.

Intra-Group Comparisons Related to Data Loss

One of the major problems in this study, as is the case with many other field research projects, is that of data loss. In this study, data were lost in two ways. First, patients dropped out of treatment prior to the three session criterion measurement. Secondly, some patients were still in treatment by the end of the study and therefore did not have the termination data completed on them. An extensive series of analyses were conducted to examine the characteristics of these groups and to compare them to subjects within the same experimental or control group on whom the hypothesis-testing analyses were conducted. This helped to ascertain how representative the data used in the hypothesis-testing comparisons were in relation to the initial subject groups admitted to the study.

A detailed presentation of these results is reported elsewhere (Stahler, 1982). In general, however, dropouts tended to be rated better on Psychological Functioning than the non-dropouts, and the non-terminated patients tended to have a higher incidence of psychosis but a lower rate of Anti-Social Behavior than terminated patients.

DISCUSSION

The purpose of this study was to assess the presence of therapist rating bias as a function of potential observation, or in this case peer review scrutiny in a field setting. It was expected that therapists would bias their ratings of client level of functioning toward greater severity so as not to "miss" any psychopathology. The results indicated partial support for the hypothesis in that the E Group patients were rated as more dysfunctional at admission than C or P Group patients on Psychological Functioning, but not in terms of the other dimensions of Basic Life Functioning, Anti-Social Behavior, or Mental Processes. This pattern of results may be because clinicians are more attuned to Psychological Functioning than to the other dimensions. The practitioner in the field of mental health carries a cognitive set predisposed to finding psychopathology. The basic purpose of an initial clinical evaluation is to assess, with as much precision and specificity as possible, the patient's symptomatology and psychological disturbance. With few exceptions, most psychotherapy orientation focuses on psychological dysfunction or maladjustment (as opposed to psychological health), just as medicine seeks to identify and treat pathological processes. Thus, it would be expected that in anticipation of possible peer scrutiny, clinicians may have a bias toward rating patients as more dysfunctional during their initial assessment since positive regard by peer professionals is in part contingent upon how well therapists can identify and recognize psychopathology. The fact that only ratings of Psychological Functioning were found to be biased, and not the other

factor scales, may be due to the fact that psychological functioning is the aspect of functioning to which clinicians are most attentive.

At termination, however, ratings of Psychological Functioning in the E Group were no different from those of the other groups. This may have been a result of the therapists' need to see their patients as improved. Given the amount of therapist effort and involvement provided in psychotherapy, clinicians may need to believe that their patients' level of functioning has improved. However, the latter findings also may have been influenced by data loss incurred in the study. Admission scores were not affected by treatment "dropout" and non-terminated clients since ratings were made during the first clinical contact. There was an approximate sample loss of 40 percent across groups between admission and termination. It is difficult to ascertain what the impact of this data loss was on the hypothesis testing. In very general terms, dropouts tended to be less dysfunctional than non-dropouts, but the non-terminated client group tended to have a higher proportion of diagnosed psychotic patients with a smaller proportion of substance abusers. Thus those with termination data utilized in the analyses may have been biased toward a more "average" level of dysfunction since there was a tendency toward a disproportionate loss of patients at opposite ends of the functioning scale.

One of the problems that lessened the impact of implementing a new problem-based quality assurance program was that the center had a peer review system already in place prior to the study. Although cases were selected purely on a random basis, clinicians were still to some extent accustomed to having their work come under peer scrutiny. Hence, the new quality assurance program involving peer review of problem cases may not have affected clinicians at this center as much as it might have at a clinic where no prior peer review had been in existence. Moreover, because research is conducted on a fairly routine basis at this particular CMHC, clinicians are perhaps more accustomed to observation than at many other clinical settings. Thus, the impact of the experimental manipulation was probably diluted somewhat by a "desensitization" to researcher scrutiny.

Another limitation of the study involved the assignment of clinicians to subject groups. Unfortunately, the same clinicians treated patients in all three subject groups. Ideally, it would have been preferable to randomly assign therapists to treat patients in each specific subject group exclusively. Although therapists were informed of subject group membership for their patients (i.e., whether these patients could be potentially selected for peer review or not), this still may have resulted in a greater probability for a Type II error. There would probably be a greater likelihood of more similar ratings of patients across groups by the same therapist, than if the ratings were made by separate therapists for each subject group. These limitations would probably increase the probability of Type II error, that is, it would be more difficult to find differences among groups. Hence, the results are probably conservative.

The major finding of the present study was that under potential observational or evaluative conditions, therapists tend to rate patients as being more psychologically dysfunctional at admission than they would otherwise rate them. This

raises the issue in psychotherapy research, program evaluation, and quality assurance systems that observing or evaluating therapist work may inherently bias therapist ratings of patient functioning. The present study was exploratory in that it was conducted in the field and examined rating bias in a rather global fashion. More controlled, rigorous research which attempts to examine more specifically the nature of therapist rating bias and isolate specific causal factors needs to be conducted. It is important to know how and in what ways the act of evaluating and studying psychotherapy itself influences that which is being studied since it is possible that our knowledge base obtained from psychotherapy research may be systematically biased in an unknown way. Research needs to address how various modalities of observation impact on the therapist—patient interactions, the extent and magnitude of the impact, and how possible consequences (such as negative peer review evaluations) affect therapist work. The often presumed independence between object of study and observer that underlies the foundation of psychotherapy research and program evaluation needs to be examined with the same scientific rigor and methods as any other psychological subject of study.

REFERENCES

Bednar, R., & Shapiro, J. Professional research commitment: A symptom or a syndrome? *Journal of Consulting and Clinical Psychology, 34,* 1970, 323-326.

Berg, I. Measures before and after therapy. *Journal of Clinical Psychology, 8,* 1952, 46-50.

Berman, L. Counter-transference and attitudes of the analysis in the therapeutic process. In M. Cohen (Ed). *Advances in Psychiatry.* New York: W.W. Norton and Co., 1959.

Carter, D.E. & Newman, F.L. *A Client-Oriented System of Mental Health Service Delivery and Program Management: A Workbook and Guide,* NIMH Series EN No. 4. Washington, D.C., GPO, 1976.

Cutler, R. Counter-transference effects in psychotherapy. *Journal of Consulting Psychology, 22,* 1958, 349-356.

Dailey, C. The effect of premature conclusion upon the acquisition of understanding a person. *Journal of Psychology, 32,* 1952, 133-152.

Edwards, M., McGuirk, F., & Wilson, N. The Fort Logan problem screen and level of functioning instrument. Paper presented at the Annual American Psychological Association Conference, Toronto, 1978.

Ellis, R. Colorado level of functioning scale. Unpublished report, Colorado Division of Mental Health Statistical Analysis and Research Section, Denver, 1977.

Ellis, R., Wilson, N., and Foster, F. Statewide treatment outcome assessment in Colorado: The Colorado Client Assessment Record (CCAR). *Community Mental Health Journal, 20,* 1984, 72-89.

Epstein, L., & Feiner, A. (Eds.) *Counter-transference.* New York: Jacob Aronson, 1979.

Page, S., & Yates, E. Fear of evaluation and reluctance to participate in research. *Professional Psychology, 5,* 1974, 400-408.

Parloff, M., Kelman, H., & Frank, J. Comfort, effectiveness, and self-awareness of criteria of improvement in psychotherapy. *American Journal of Psychiatry, 111,* 1954, 343-352.

Reich, A. On counter-transference. *International Journal of Psychoanalysis, 32,* 1951, 25-31.

Rosenberg, M. The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. Rosnow (Eds.) *Artifact in Behavioral Research.* New York: Academic Press, 1969, 279-349.

Rosenthal, R. *Experimental Artifacts in Behavioral Research.* New York: Appleton-Century-Crofts, 1966.

Rosenthal, R., & Rosnow, R. (Eds., *Artifact in Behavioral Research.* New York: Academic Press, 1969.

Searles, H. *Counter-transference and Related Subjects.* New York: International Universities, 1979.

Stahler, G. *An Assessment of Therapist Rating Bias and the Hawthorne Effect in a Program Evaluation.* Doctoral dissertation, Department of Psychology, Temple University, 1982.

Steinhalber, J. Bias in the assessment of psychotherapy. *Journal of Consulting and Clinical Psychology, 34,* 1970, 37-42.

Sullivan, H. *Personal Psychopathology.* Washington, D.C.: William Alanson White Psychiatric Foundation, 1965 (unpublished).

Tash, W., and Stahler G., The history and current status of mental health quality assurance. *American Behavioral Scientist,* Spring, 1984.

Tash, W., Stahler, G., & Rappaport, H. Evaluating quality assurance programs. In G. Stahler and W. Tash (Eds) *Innovative Approaches to Mental Health Evaluation.* New York: Academic Press, 1982, 113-138.

Ward, C. and Richards, J. Psychotherapy research: inertia, recruitment, and national policy. *American Journal of Psychiatry, 124,* 1968, 1712-1714.