Assessment of Outcome by Independent Clinical Evaluators

Robert L. Spitzer, M.D.* and Jean Endicott, Ph.D.*

General Methodological Issues

For research purposes, there are many advantages to having an assessment of treatment outcome by a clinician who has not been involved in the treatment being evaluated. An independent assessment avoids possible bias in evaluations made by the therapist or other members of the treatment team, such as nurses or attendants. Of course, patient self-report may avoid therapist bias but the patient may himself be inclined towards either overestimating or underestimating the effect of the treatment he has received. In addition, patient self-evaluation is necessarily limited to areas which the patient can judge, such as his mood, attitudes, and some obvious aspects of functioning. Patients often cannot evaluate their ability to test reality accurately or be aware of impairment in their thought processes. These areas of functioning usually require judgment by another person.

An additional advantage of an independent evaluation is that it is usually awkward for a clinician to make a systematic, comprehensive, cross-sectional assessment of a patient he has been treating for some period of time. While he is usually able to give valid ratings of those aspects of the patient's functioning which have been the subject of his interaction with the patient, he will often be tempted to make assumptions about other areas of functioning rather than ask about them directly. The research interviewer who explains to the patient the need for an independent assessment does not face the awkward situation of the therapist, who is reluctant to ask about an area of behavior (such as hallucinations or sexual perversion) which he knows is almost certainly not present.

Like everything else in life, an independent clinical assessment poses potential problems. Some patients do not readily accept the idea of being interviewed by a stranger when

*Dr. Spitzer is Director, and Dr. Endicott is Co-Director, Evaluation Section, Biometrics Research, New York State Department of Mental Hygiene at the New York State Psychiatric Institute.
the purpose of the interview is clearly not related to helping them. However, our experience indicates that if patients are told by both their therapists and the research interviewers that it is often helpful to get another view of the patient’s condition, most patients will cooperate.

More frequently the therapist — not the patient — is reluctant to have the results of his work with the patient examined by a colleague. This reluctance may be rationalized by the argument that such an assessment interferes with the ongoing therapeutic process. Some therapists, on the other hand, have found that the patient’s reaction to such a procedure can be usefully dealt with in therapy.

The most serious problem raised by an independent clinical assessment is the issue of validity: can one trust judgments made on the basis of contact which is necessarily limited to a short time period of one or two hours. There is no doubt that some patients are capable of presenting themselves, either deliberately or unknowingly, in a manner which does not correspond to their usual behavior. However, evidence from our own work and from the work of Lorr, Overall, Gorham and many others, demonstrates that for purposes of group comparison independent assessments provide valid ratings of patient behavior.

The use of an independent clinical evaluation is most useful when the purpose of the assessment is to ascertain the subject’s level of overt symptomatology and the more obvious aspects of impaired functioning. More subtle personality features, such as defensive patterns or the quality of interpersonal relationships, are more difficult to evaluate and generally require more familiarity with the subject than is afforded by an independent evaluation. However, it should be noted, that these more subtle areas are difficult to evaluate reliably under any circumstances.

In choosing a specific evaluating instrument, the independent assessor must consider several factors.
1. Coverage. Although an individual investigator may wish to focus on a limited and specific area of interest, such as interpersonal relationships, there are advantages to having an instrument which includes coverage of aspects that interest most investigators in the field. There is clearly most consensus about the need for evaluating overt symptomatology and impairment in functioning. Although many psychotherapy investigators, particularly those with a psychoanalytic orientation, are more concerned with intrapsychic functioning, most clinicians would agree that important changes in intrapsychic functioning would necessarily be accompanied by changes in overt symptomatology.

Another advantage to including coverage of overt symptomatology is that the behaviors of interest can be described in such a way as to avoid theoretical assumptions. The evaluating instrument can therefore be used to compare therapeutic approaches which have different theoretical assumptions.

Many rating scales have proven to be extremely valuable in the assessment of psychopathology, particularly for psychopharmacology research. However, most of the existing instruments do not assess alcoholism, drug addiction, and antisocial behavior despite the fact that individuals with disturbances in these areas represent a significant portion of the patients now being seen at community mental health centers and by private practitioners. Moreover, very few instruments assess impairment in functioning in the roles of employed wage earner, housekeeper, student, parent or mate. Likewise, few instruments cover impairment in leisure time or daily routine activities. Using instruments which do not cover these areas may result in a failure to detect important differences in therapeutic effectiveness.

The inclusion of coverage of historical information relevant to prognosis, severity of illness and diagnosis would markedly improve the value of an independent assessment made at the beginning of treatment. Most studies of prognosis have indicated that historical and demographic information
are often far more predictive of outcome than a cross-sectional evaluation of symptomatology at the time of beginning of therapy.

2. Length of time to administer. All investigators perpetually search for a five minute interview form of proven reliability and validity which includes comprehensive coverage of all possible areas of interest to them. Unfortunately, a more realistic appraisal suggests that an independent assessment often requires at least an hour if the evaluation is at all comprehensive. However, when the interview takes more than two hours, increasing difficulty results because of subject and interviewer fatigue and annoyance.

3. How to obtain the information. Information from the patient can be obtained with the usual clinical interview. However, during the past few years attempts have been made to improve the research value of a psychiatric interview by standardizing the interview techniques so that the variability associated with differences in interviewing methods and coverage is reduced. Interview schedules have been developed which combine the flexibility and rapport that are inherent in a clinical interview with the completeness of coverage and comparability of interviewing method that result from using a structured interview procedure.

4. Personnel. The decision about the level of training required of the independent assessor is often determined by the practical considerations of money and availability of highly-trained personnel, such as psychiatrists. By using an interview schedule and precoded items which are at a descriptive level that avoid the use of technical terms, we have found that interviewers who are not psychiatrists or clinical psychologists can be trained to make and record accurate judgments about the psychopathology of individuals they interview.

5. Psychometric properties. If possible, an investigator should use an instrument which has demonstrated reliability and validity. Although there are many kinds of reliability, in
using an independent assessment procedure one is generally concerned with inter-judge agreement across a series of patients similar to the patients who are to be studied. There are also several kinds of validity. In psychotherapy research the instrument is most often used as a measure of current state prior to treatment and as a criterion measure at various follow-up periods. Therefore, the investigator should know what evidence there is for concurrent validity of the instrument as a measure of the dimensions in which he is interested. Instruments are often chosen because they are popular, machine-scoreable or simple to administer even though they may not be measuring the dimensions appropriate to the hypotheses being tested in the research study.

Recommendations of Specific Instruments

We recommend four instruments for consideration in designing psychotherapy outcome studies. The reader should know that we are not completely unbiased: we and our colleagues developed three of these four. These three are: the Psychiatric Status Schedule (PSS), the Psychiatric Evaluation Form (PEF), and the Current and Past Psychopathology Scales (CAPPs). These three instruments are essentially alternatives to each other, and in any given study it is unlikely that more than one of them would be used. The fourth is the Health Sickness Rating Scale (HSRS) which was developed by Lester Luborsky and his colleagues at the Psychotherapy Research Project of the Menninger Foundation.

Psychiatric Status Schedule, Psychiatric Evaluation Form, and Current and Past Psychopathology Scales

The PSS, the PEF and the CAPPs were designed to improve the research value of clinical judgments of psychopathology and role functioning based on data collected during a psychiatric interview. Each of them has been described in detail in a separate paper, and their
major features are contrasted in Table 1. All three instruments use an interview schedule which the interviewer uses to elicit information needed to judge the precoded items or scales. The interview schedules provide a natural progression of topics and consist of a series of statements and predominantly open-ended questions. When skillfully administered, they produce the flavor of a clinical interview. The interviewer is encouraged to use supplementary questions to clarify or probe areas about which he may need more information to make the necessary judgments.

1. Coverage: These three instruments contain sections which evaluate the usual mental status type of signs and symptoms of psychiatric disorder as well as 1) impairment in formal role functioning; 2) impairment in the efficiency and conduct of leisure time and daily routine activities; 3) impairment in interpersonal relationships; 4) the use of drugs and alcohol; and 5) other illegal or antisocial activities. The focus of the PSS and PEF is upon the subject's symptoms and functioning during the week prior to the interview, except for those sections dealing with role functioning and the use of alcohol and drugs, where the time period covered is one month.

The coverage of the CAPPS is divided into two parts: the current section focuses upon the last month, and the past section deals with the period between age 12 and up-to-the-last-month.

The interview schedules of each instrument are accompanied by inventories of carefully defined items or rating scales which use non-technical descriptions of behavior. All of the data are obtained, judged, and recorded during the interview so that the interview and evaluation are completed simultaneously.

In the PSS, the judgments are recorded in 52 dichotomous items describing small units of behavior. In the PEF and the CAPPS most of the judgments are recorded on six-point scales which describe broad dimensions of behavior. Many of these dimensions correspond to factors which emerged from factor analytic studies of the PSS items. Thus the PSS gives more specific details
about a subject’s functioning while the PEF and CAPPS allow the interviewer to make summary judgments of large dimensions, taking into account both the intensity and duration of the different elements which make up the dimension.

2. Populations for which instruments are relevant. These instruments were constructed to make the coverage relevant to an extremely wide range of psychiatric disorders. Although there are items or scales appropriate to psychotic patients likely to be inpatients, there is also extensive coverage of less severe psychopathology likely to be present in an outpatient group. The three instruments have also been utilized with individuals who have never been in psychiatric treatment.

3. Feasibility. Because they provide interview schedules and information about the nature of the judgments to be made, these three instruments can be used effectively by interviewers who have not had a great deal of clinical experience and training. Since the PSS contains descriptions of small units of behavior, and thus largely avoids the problem of scaling of severity, even less clinical experience with a wide range of patients is required of the interviewer who uses this form. Training materials and suggested procedures for training interviewers are available for all three instruments. Psychiatrists and clinical psychologists probably require only a few hours of training, but more time and personal instruction are needed by individuals with less experience, such as college students, research assistants, nurses, social workers and medical students.

Using a structured interview facilities comparability of coverage across interviewers and makes it possible to survey a wide range of behavior in a relatively short period of time. The PSS, the PEF and the current section of the CAPPS each take about 30 to 50 minutes. The past section of the CAPPS takes an additional 30 to 50 minutes.

In successively revising the instruments we eliminated or reworded questions which might be offensive to some subjects. The interviewing procedures for the three instruments are
sufficiently brief so that there is rarely an objective to the length of the interview. Moreover, when the instruments are properly explained to the subject, the fact that they focus on psychopathology has not proven to be a barrier to easy communication. Many subjects have commented favorably on the thoroughness of the interviews.

4. Psychometric characteristics. Each instrument has been subjected to factor analytic studies which have resulted in the development of factor based summary scoring systems. The names of the factor based scales of the three instruments are listed in Table 2.

The interjudge reliabilities for the three instruments are very high. These instruments have been shown to discriminate among different diagnostic groups and groups known to differ in severity of illness. Several studies have demonstrated that they are sensitive to change. In addition, each instrument has proven to have significant correlations with other instruments, such as the MMPI and other self-report measures which purport to measure similar dimensions.

A special feature of both the PSS and the CAPPS is found in the availability of a computer program for psychiatric diagnosis. The PSS provides input for DIAGNO I which yields one of 27 standard psychiatric diagnoses including Not III. The CAPPS provides input for DIAGNO II which yields one or more of 46 psychiatric diagnoses. Several studies have demonstrated the usefulness of the computer diagnosis as an overall summary description of the patient's status at the time of evaluation.

Interested readers will find a full description of reliability and validity data in several papers (1-7).

5. Data Processing. Computer programs are available for scoring each of the instruments and for making grouped comparisons of individual items or scales or summary scores.

6. Relationship to Theoretical Formulations Regarding Psychotherapeutic Change.
Since the items and scales of the PSS, PEF and CAPPS are deliberately formulated in simple descriptive language, avoid the use of terminology specific to any one theoretical framework, focus on manifest behavior, and are broad in their coverage, these instruments can be used to compare varied forms of psychotherapy despite their different aims and theoretical formulations.

Each instrument covers areas of functioning that can be expected to change in patients treated psychotherapeutically regardless of the specific type of therapy. Even when the therapeutic efforts are directed towards specific units of behavior, as might be the case with behavior-oriented therapy, one would expect that successful treatment would create some "ripple effect" that would produce improvement in some broad dimensions of behavior measured by one of the factor scales.

Relative Advantages of the Three Instruments

Each of the three instruments has certain advantages over the other two. The PSS can be administered by personnel who have had less clinical experience and training than that required for the other instruments since the judgments are dichotomized and do not involve an assessment of severity. The PEF interview is usually shorter than a PSS or CAPPS interview. The main advantage of the CAPPS is the inclusion of historical material which may be of relevance to the prediction of differential response to psychotherapy, and for describing the sample more thoroughly. For follow-up evaluations, only the current section of the CAPPS need be repeated. For some investigators, the PSS would have an advantage in that the judgments are recorded in items describing smaller units of behavior. This permits the construction of new scales combining various subsets of items.

Menninger Health Sickness Rating Scale

We recommend that the Menninger Health Sickness Rating Scale (HSRS) be used in conjunction with one of the three structured interviews discussed above. This instrument is a global
assessment of severity of illness that uses a zero to 100 scale with anchor points that are described in
diagnostic and functional terms. In addition, the evaluator has available brief descriptions of actual
patients whose scores are distributed throughout the range of the scale. When utilizing this scale, the
clinician takes into account the patient’s need to be protected or supported by the therapist or
hospital, the seriousness of symptomatology, the degree of subjective discomfort, the effect on his
environment, the degree to which he can make use of his abilities, the quality of interpersonal
relationships and the breadth and depth of his interests. One of the HSRS’s major advantages is that
the rater can take into account information that may not be recorded in ratings of specific dimensions.

We have used this instrument in a number of studies of inpatients, outpatients and
nonpatients. Many of our comments below are based on this experience.

1. Psychometric information. Interjudge reliability is quite high. The concurrent validity
of the HSRS has been demonstrated by sizable correlations with other measures evaluating overall
severity of illness. As would be expected, it is sensitive to change in the patient’s level of functioning
and severity of symptomatology. Detailed psychometric information is available and is noted in Dr.
Dr. Luborsky’s discussion of the HSRS.

2. Population. The instrument is designed for use with the entire spectrum of
severity of illness. However, we have found that raters sometime have difficulty in rating certain kinds
of patients. Since the evaluator is expected to take into account both underlying illness and level
of functioning, he often does not know how to rate a schizophrenic who is in a period of remission and
virtually asymptomatic. Similarly, it is sometimes difficult to rate a patient whose behavior is so
erratic that during the two week suggested time period he may have been acutely suicidal for a few
days and relatively asymptomatic the rest of the time. It is not clear if the rater is to take the lowest
level of functioning or make a weighted average.
3. Feasibility. The HSRS is easy to use; once the rater is familiar with the anchor points the actual rating takes no more than a minute of reflection.

4. Personnel. The HSRS would seem to be appropriate only for raters who have considerable experience with the full range of psychiatric disability.

5. Relationship to theoretical formulations. The seven criteria previously noted that go into the severity ratings are atheoretical; few clinicians would question their relevance to severity of illness. However, those investigators who have little or no use for the concept of diagnosis may object to its inclusion in the examples of the anchor points and descriptions of the sample cases.
References


<table>
<thead>
<tr>
<th>Instrument</th>
<th>Areas Covered</th>
<th>Time Period</th>
<th>Judgments</th>
<th>Time to Administer or Complete</th>
<th>Physical Layout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychiatric Status Schedule (PSS)</td>
<td>Neurotic, organic and psychotic signs and symptoms plus impairment in daily routine, use of leisure time and travel plus alcoholic, drug addictive and psychopathic behavior. Role functioning as wage earner, housekeeper, student, mate, and parent</td>
<td>Last week</td>
<td>321 dichotomous and check list items</td>
<td>30-50 minutes</td>
<td>21 page step-down booklet and scoresheets</td>
</tr>
<tr>
<td>Psychiatric Evaluation Form (PEF)</td>
<td>Same as PSS plus characteristics of present illness and major reason for admission</td>
<td>Last week</td>
<td>27 scales and 2 check list items</td>
<td>With interview guide, 30-50 minutes</td>
<td>9 page step-down booklet and scoresheets</td>
</tr>
<tr>
<td>Current and Past Psychopathology Scales (CAPPS)</td>
<td>Similar to PEF plus history relevant to severity, prognosis and diagnosis</td>
<td>Last month and age 12 to last month</td>
<td>171 scales and check list items</td>
<td>With interview guide, 1-2 hours</td>
<td>13 page protocol (with interview schedule 26 pages)</td>
</tr>
</tbody>
</table>
### Table 2

**Factor-based Scales of the Psychiatric Status Schedule, Psychiatric Evaluation Form and the Current and Past Psychopathology Scales**

<table>
<thead>
<tr>
<th>Psychiatric Status Schedule</th>
<th>Psychiatric Evaluation Form</th>
<th>Current and Past Psychopathology Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symptom Scales</strong></td>
<td><strong>Factor Scales</strong></td>
<td><strong>Current Scales</strong></td>
</tr>
<tr>
<td>Depression—Anxiety</td>
<td>Disorganization</td>
<td>Reality Testing—Social Disturbance</td>
</tr>
<tr>
<td>Daily Routine—Leisure Time</td>
<td>Subjective Distress</td>
<td>Depression—Anxiety</td>
</tr>
<tr>
<td>Impairment</td>
<td>Antisocial</td>
<td>Impulse Control</td>
</tr>
<tr>
<td>Social Isolation</td>
<td>Withdrawal</td>
<td>Somatic Concern—Functioning</td>
</tr>
<tr>
<td>Suicide—Self Mutilation</td>
<td>Alcohol</td>
<td>Disorganization</td>
</tr>
<tr>
<td>Speech Disorganization</td>
<td>Grandiosity—Externalization</td>
<td>Obsessive—Guilt—Phobic</td>
</tr>
<tr>
<td>Inappropriate Affect,</td>
<td></td>
<td>Elation—Grandiosity</td>
</tr>
<tr>
<td>Appearance or Behavior</td>
<td>Role Scales</td>
<td>Summary Role</td>
</tr>
<tr>
<td>Agitation—Excitement</td>
<td>Denial of Illness</td>
<td>Past Scales</td>
</tr>
<tr>
<td>Interview Belligerence—</td>
<td>Wage Earner Role</td>
<td>Depression—Anxiety</td>
</tr>
<tr>
<td>Negativism</td>
<td>Housekeeper Role</td>
<td>Impulse Control</td>
</tr>
<tr>
<td>Disorientation—Memory</td>
<td>Student or Trainee Role</td>
<td>Somatic Concern—Functioning</td>
</tr>
<tr>
<td>Retardation—Lack of Emotion</td>
<td>Mate Role</td>
<td>Obsessive—Compulsive</td>
</tr>
<tr>
<td>Antisocial Impulses or Acts</td>
<td>Parent Role</td>
<td>Anger—Excitability</td>
</tr>
<tr>
<td>Drug Abuse</td>
<td></td>
<td>Manic</td>
</tr>
<tr>
<td>Reported Overt Anger</td>
<td></td>
<td>Sexual Disturbance</td>
</tr>
<tr>
<td>Grandiosity</td>
<td></td>
<td>Memory—Orientation</td>
</tr>
<tr>
<td>Suspicion—Persecution—</td>
<td></td>
<td>Disorganized</td>
</tr>
<tr>
<td>Hallucinations</td>
<td></td>
<td>Organicity</td>
</tr>
<tr>
<td>Alcohol Abuse</td>
<td></td>
<td>Neurotic Childhood</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phobia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Retardation—Stubborn</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hysterical Symptoms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intellectual Performance</td>
</tr>
</tbody>
</table>

*The six factor scales of the Psychiatric Evaluation Form are the result of the factor analyses of its 19 symptom scales. These symptom scales similar to the symptom scales of the Psychiatric Status Schedule.*