Documenting an Episode of Psychiatric Illness: Need for Multiple Information Sources, Multiple Raters, and Narrative

by Ian F. Brockington and Herbert Y. Meltzer

Abstract

It is argued that the structured psychiatric interview, which has become the foundation of much modern clinical research, is not a sufficient method of documenting an episode of psychotic illness. This valuable resource needs to be supplemented by a high standard of recorded observation throughout hospitalization by all staff in contact with the patient. The conversion of observations into ratings can involve considerable error and should be left to at least two highly trained research workers. The only data worthy of this effort are the entire body of clinical information obtained from all sources. This requires detailed written records. The advantage of a narrative record, in addition to numerical ratings, is that it can be reinterpreted at any time using different systems of psychopathological or nosological ideas.

Documenting the signs and symptoms of a psychotic episode is often difficult because of the nature of the material. The subject matter consists of mental experiences and altered behavior—phenomena that are inherently difficult to comprehend, measure, and describe. When ascertaining the symptoms, the observer is at the mercy of the patient's own perceptual and descriptive limitations, aggravated by mental illness, while behavioral events are transitory, unpredictable, and hard to scrutinize. Clinical data, therefore, are liable to be riddled with inaccuracy, greatly increasing the difficulty of all research which involves dividing patients into groups, or measuring the change of state. In some disciplines, such as epidemiology, it is possible to compensate for observational error by large numbers, but it is often necessary to manage with small samples, especially now that psychiatrists are narrowing their focus and trying to draw finer distinctions between psychotic states. For such studies it is important to develop clinical techniques which are precise and sensitive, documenting the phenomena with minimum error. This article addresses the questions of how clinical psychiatric research should be conducted in order to obtain a more complete and accurate picture of a psychotic episode.

The Need for Multiple Information Sources

The Structured Psychiatric Interview. The structured interview is the mainstay of modern clinical study. The principles of interviewing are well understood. The interviewer prepares by reading the case records before seeing the patient, forms a relationship with him explaining the purpose of the interview, allows time for his spontaneous complaints before proceeding to direct questioning, uses the interview flexibly but asks all the recommended probes, avoids leading questions, cross-examines and clarifies the nature of the symptoms, and never accepts the presence of a phenomenon unless it is clearly described in the patient's own words. After the interview, he completes a set of ratings aided by a glossary. This

Reprint requests should be sent to Dr. I. F. Brockington at Department of Psychiatry, University Hospital of South Manchester, West Didsbury, Manchester M20 8LR, United Kingdom.
form of inquiry has been developed to a high level at the New York State Psychiatric Institute (Spitzer et al. 1970) and the M.R.C. Social Psychiatry Unit in London (Wing, Cooper, and Sartorius 1974), but there is a danger that its use will lead to complacency and overconfidence, giving investigators a false sense of the thoroughness, reliability, and validity of their clinical assessment. It should provide a comprehensive inventory of recent symptoms provided that the patient is capable of giving a clear account—a condition not met in about 25 percent of psychoses (Brockington and Leff 1979). The behavioral observations are less likely to be representative of the whole episode because of the brevity of the contact between observer and patient. When the aim is to document an episode lasting weeks or months, a single 1-hour interview, however thorough the observer's methodology and training, is not a sufficient source, and its limitations have been shown by measurements of its test-retest reliability and validity. Some test-retest reliabilities, taken from studies of the 7th and 8th editions of the Present State Examination (Wing et al. 1967; Kendell et al. 1968), the Psychiatric Status Schedule (Spitzer et al. 1970), the Schedule for Affective Disorders and Schizophrenia (Endicott and Spitzer 1978), and the International Pilot Study of Schizophrenia (IPSS) (World Health Organization 1973), are shown in table 1. Reliability improves as one moves from items to section scores and syndromes. At the level of syndromes—called "groups of units of analysis" in the IPSS—or diagnoses, it is good, but for items it is poor. The reasons for the discrepancies are unclear. They may be due to variations in interview administration or scoring which could be removed by further standardization and training. Alternatively they may be due to variation in the patient's description of his symptoms; he may be reporting his state of mind on a particular day or in the immediate past and not during the last month or the whole episode. Whatever the reason, the test-retest reliability measurements suggest that only the broad outlines of an episode are agreed upon in two interviews held a few days apart. When the aim is to find distinctions between psychotic states, it is not likely that broad outlines will provide the answer. For example, the distinction between bipolar and other depressive illnesses at present depends mainly on the occurrence of manic episodes, because we cannot reliably distinguish manic-depressive disease in its depressed form from unipolar depressive disorder. When searching for psychopathological differences between bipolar and nonbipolar depression, it is probable that detailed symptomatology—for example, the severity of retardation, the worsening of symptoms in the morning, the lack of anxiety, and the presence of particular kinds of delusion—will prove discriminating, rather than overall section scores. It is necessary, therefore, to achieve satisfactory reliability at the item level.

There have, to our knowledge, been only two attempts to study the validity of the psychiatric interview by comparing it with a criterion that was at least partly independent and based on a larger body of data. Carpenter et al. (1976) compared interview ratings with those made by the psychiatrist-in-charge after a month's observation. The interview ratings were unseen, but the clinician attended the research interview. They found that the interview missed 30 percent of the psychopathology which later became evident. Downing, Francis, and Brockington (1980) compared five information sources with a criterion. The sources comprised a psychiatric interview...
chiatric interview, interview with a relative, nurse rating, self-rating and brief videotaped interview.

The criterion was a set of "master ratings" made by three observers reviewing all the research data (except self-ratings) and the hospital records. It was independent of the self-ratings but not independent of the other sources, which contributed to varying extents. The psychiatric interview contributed about 20 percent (800 words) of the information available, but in terms of detailed verbatim replies to key questions, its contribution was greater. In spite of this, it missed over half the items agreed to be present, and there was an overall underrating of the severity of phenomena. Some of its deficiencies are listed in tables 2 and 3.

Multiple Research Interviews. To supplement the psychiatric interview, the first resource is the use of several interviews. It is good practice to conduct the initial interview in two stages. The first is done shortly after admission and has the purpose of establishing rapport, observing the patient's state when most severely disturbed, recording his complaints, and checking his cognitive function. The second is done a day or two later to make a second set of behavioral observations, assess the persistence of the main symptoms, and complete the routine probes. Two shorter interviews make fewer demands on both patient and interviewer. A psychiatric interview is demanding because the observer is involved simultaneously in several different roles: He has to maintain a professional relationship with the patient, treating him with respect and concern, and giving him scope to express his preoccupations; he has to control the dialogue and ensure that every area is explored; he has to evaluate the symptomatology, matching the patient's description against the glossary; and he has to switch into an objective stance when rating affect, speech, and behavior.

In addition to the initial interviews, it seems of value to conduct a full structured interview at the time of discharge. The patient may then be able to describe experi-

| Table 2. Some low concordances between psychiatric ratings and master ratings |
|-----------------|-----------------|-----------------|-----------------|
| Items           | Item rated in psychiatric interview | Items agreed to be present | Kappa |
| Agitation       | 3               | 20              | .14             |
| Excitement and violence | 2               | 14              | .19             |
| Perplexity      | 2               | 12              | .23             |
| Incongruous affect | 6               | 17              | .38             |
| Lability of mood | 5               | 15              | .41             |
| Guilt           | 10              | 23              | .41             |
| Visual hallucinations | 4               | 12              | .43             |
| Denial of abnormality | 10              | 21              | .43             |
| Panic attacks   | 3               | 8               | .44             |

Table 3. Some low correlations between psychiatric ratings and master ratings

<table>
<thead>
<tr>
<th>Scales</th>
<th>Mean score from psychiatric interview</th>
<th>Mean score from master ratings</th>
<th>Pearson's r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concern for others</td>
<td>18</td>
<td>30</td>
<td>.03</td>
</tr>
<tr>
<td>Incompetence</td>
<td>21</td>
<td>44</td>
<td>.05</td>
</tr>
<tr>
<td>Social withdrawal</td>
<td>17</td>
<td>29</td>
<td>.23</td>
</tr>
<tr>
<td>Manipulative behavior</td>
<td>8</td>
<td>9</td>
<td>.23</td>
</tr>
<tr>
<td>Uncooperative behavior</td>
<td>13</td>
<td>30</td>
<td>.31</td>
</tr>
<tr>
<td>Self-neglect</td>
<td>10</td>
<td>13</td>
<td>.49</td>
</tr>
<tr>
<td>Confusion</td>
<td>7</td>
<td>18</td>
<td>.51</td>
</tr>
<tr>
<td>Worrying</td>
<td>20</td>
<td>33</td>
<td>.53</td>
</tr>
<tr>
<td>Somatic concern</td>
<td>11</td>
<td>20</td>
<td>.59</td>
</tr>
</tbody>
</table>

*p<.05.  
**p<.01.
ences still recent and vividly remembered which he could not clearly describe on admission because of guarding, muteness, or incoherence. The time of discharge is an opportunity to assess the degree of recovery, which is considered one of the most important variables in making a diagnosis. Once the patient has left the hospital it may be difficult to trace and interview him again.

Case Records. Research records can be supplemented by interviews recorded in the case notes (charts). These are held throughout the period of hospitalization and record the course of the episode. The information has been gathered without the discipline of the formal research interview, but is not thereby invalidated. All trainee psychiatrists can be trained in the method of interviewing, which is simpler than training them in the practice of rating. If they are taught to record phenomena in plain language rather than professional jargon—to refrain from converting their observations into inferential terms whose definitions vary from time to time and from center to center (e.g., delusions, thought disorder, blunting of affect) and report what was said and what they saw—their information would have the same value as the research interview, lacking only the range of the full inventory. The limitations would be set by their personal observational and writing skills rather than the formality of the method. In some centers (e.g., the Professorial Unit at the Maudsley Hospital) a high standard of recordkeeping has been maintained for many years, and it is possible to use these data for nosological research (e.g., Cutting, Clare, and Mann 1978). In the study of the longitudinal course of psychotic illness, the general quality of hospital notes is a limiting factor since few of the episodes can be studied by a research team.

Interviews With Other Informants. So far this article has addressed what can be learned by interviewing the patient. The particular value of interviewing a close relative or cohabitee is that his frame of reference includes the patient’s normal personality and he can report the contrast between it and the present state. He can give an objective account of the historic development of the episode and of major events preceding admission such as suicide attempts. Relatives can be interviewed in a high percentage of cases—88 percent in the study of Downing, Francis, and Brockington (1980). One can use a structured interview similar to that used to interview the patient, with slightly modified probes. Downing, Francis, and Brockington (1980) found highly significant correlations between ratings made from this interview and master ratings in 25/57 areas studied. There were only five areas in which they were much better than those based on the psychiatric interview—social withdrawal, increased sociability, incompetence, euphoria, and increased energy. Three of these are useful in the diagnosis of mania. Like the psychiatric interview this source underestimated the severity of psychopathology in every area except anxiety and poverty of contact, and correctly identified only 38 percent of the items agreed to be present.

Nurse Ratings. After the patient’s hospital admission, members of the nursing staff take over the relatives’ role in relating to the patient in almost all the circumstances of his daily life. As Bunney and Hamburg (1963) have pointed out, nurses are in a unique position, having a well-accepted place on the unit, having direct contact with the patient in the course of their regular duties, and being able to make a continuous series of observations throughout the day and night. They observe the patient in the washroom, at meals, and relating to other patients, to staff of all kinds, and to relatives, in group meetings, and (especially if occupational therapists are drawn in) when engaged in a wide range of activities. They are able to assess the variability or persistence of affective symptoms and the frequency of abnormal behaviors, and can provide evidence for an assessment of orientation and volition. To varying extents they lack a precise glossary of terms, but they do not lack experience of psychotic phenomena. Their evidence is indispensable and is the only source always available once the patient has been hospitalized.

Downing, Francis, and Brockington (1980) studied the validity of nurses’ ratings, compared with the comprehensive criterion described above. They found that ratings made after a single 8-hour shift correctly identified only 24 percent of the items agreed to be present during the whole episode, but that nurses’ ratings were in agreement with master ratings at the 1 percent level of significance in 29/57 areas studied. The highest correlations were all concerned with behavior (retardation, loss of reserve, poverty of contact, social withdrawal, concern for others,
disorganized communication, incompetence, and uncooperative behavior with Pearson’s *r* in the range .58–.64), but there were fairly high correlations in some symptom areas too, including guilt (*r* = .58), hypochondriasis (*r* = .57), delusions (*r* = .53), and auditory hallucinations (*r* = .54).

In four areas nurses’ ratings had the best agreement with the criterion of all the sources studied—namely, uncooperative behavior, concern for others, agitation, and distractibility.

Since the introduction of ward behavior charts at the Phipps Clinic in Baltimore (Kempf 1915), a great many nurses’ rating scales and schedules have been published. One of the main problems is that nurses are less highly trained than psychiatrists and psychologists in rating psychopathology and some reports on their reliability have been adverse (Jensen and Morris 1960; Downing and Brockington 1978). In the writers’ view, the advantages of nursing observations are lost if they are all converted to ratings. The virtue of the nursing input is that it comes from a multitude of personnel working continuously with the patient. One cannot expect them all to maintain a high standard of rating, but one can expect them to report what happens. They are already accustomed to writing regular transfer notes, which provide useful data on day-to-day change and salient incidents. For research purposes it is necessary to train nurses to describe their observations in a more detailed and graphic form. The process of sharpening observational skills and improving communication will have a beneficial effect on nursing, as well as furnishing a prime source of information.

Self-Report. In theory the patient’s own account of his psychosis has a special interest and value. In practice few patients provide an unstructured account of their experiences and ideas, and a good psychiatric interview records the main symptoms verbatim. Self-diagnosis and self-ratings made by the patient have many disadvantages; the patient has no frame of reference outside his own experience, and his answers may be distorted by dissimulation or other forms of response bias. In nosological studies, however, self-rating has the advantage of being free from the observers’ theoretical prejudices, whose effects on perception, reporting, and rating are hard to eliminate. In the patient’s own ratings we have a source completely purged of this subtle and pervasive distortion.

In the Manchester study (Downing, Francis, and Brockington 1980), some observations on self-rating were made using a set of 100 questions about the current mental state to be answered “yes” or “no” by forced choice. Forty-six out of 50 patients were able to complete this procedure. From the replies, 22 scales were arbitrarily constructed, corresponding to the master rating scales. It was found that eight of them had significant correlations (p = <.001) with the master ratings. They comprised elation (*r* = .61), hopelessness (*r* = .54), guilt (*r* = .52), auditory hallucinations (*r* = .51), grandiosity (*r* = .50), social withdrawal (*r* = .49), insomnia (*r* = .46), and sadness (*r* = .44). The correlations were modest, but the criterion was entirely independent. When used in an empirical study of the difference between puerperal and nonpuerperal depression (Brockington et al., unpublished data), the self-ratings yielded an unexpected and valuable finding completely missed by observer ratings. It was not possible by observer ratings to distinguish between the severity of depression in the two groups, but self-ratings of sadness showed a marked difference, with puerperal depressions having significantly lower scores (p = .005). Thus in a broad approach to the study of mental state, self-ratings may have a limited but useful contribution to make.

Recorded Interviews. The development of videotape recordings has made it possible to make a permanent replica of interactions with patients, which has many advantages over the written record. The interview can be studied repeatedly by the most experienced raters, and difficult judgments about affect (e.g., perplexity and blunting) can be made by experts. It can be studied at leisure by observers concentrating entirely on rating without having to attend to the dynamics of the interview. In treatment trials the ratings can be made blind to time. The verbal component can be used for analysis of content and thought disorder. Above all, it is possible to make measurements, as opposed to ordinal ratings; for example, one can measure the latency of response (initial pause time), which Ahmed and Brockington (unpublished observations) have shown to have high test-retest reliability (*r* = .9994) and a high correlation with observer ratings of retardation (*r* = .79). Contrary to expectation, a videotaped interview is well tolerated by psychotic pa-
tients; in the Manchester study (Downing, Francis, and Brockington 1980), 48 of 50 patients had a 5- to 10-minute recorded interview shortly after admission. The disadvantages of the method are that a very brief sample of an episode can be recorded and only under artificial conditions; moreover, the analysis is extremely time-consuming.

Perhaps for this reason little work has so far been done on the use of the videotaped interview in clinical or nosological studies. Katz and Itil (1974) have shown that videotaped ratings were far more sensitive to changes produced by thioridazine and thiothixine in elderly patients than interview ratings using the Brief Psychiatric Rating Scales. One could envisage an expanded role for this investigatory tool. One or more 15- to 20-minute interviews during the most disturbed phase could be compared to a similar interview at the time of discharge. They would be focused on the patient’s present concerns and recent events, providing material for an analysis of verbal content, measurements of speech, and ratings of affect. They would take over some of the functions of the structured psychiatric interview, which could then be used for a deeper exploration of particular areas. For example, it would be possible to devote more time to the analysis of delusions, not simply identifying the presence of ideas of grandeur or persecution, for example, but exploring their origin, structure, and extent. This might lead to the development of better measures of delusional thinking which would help in the differentiation of paranoid psychoses from other forms of schizophrenia.

An Illustrative Case. The following case illustrates how data from several sources contribute to the full picture of a psychotic episode:

The patient was a 30-year-old woman repatriated from Africa, where she suddenly became disturbed while living in a small American enclave which a hostile government wanted to close down. When formally interviewed the day after admission, she was lethargic (heavily sedated) and replied evasively. She gave conflicting accounts of having excessive energy (e.g., gardening in the early morning) and diminished energy. She said she was being watched and whispered the words “Russians,” “persecution,” “airplanes and things,” and “making noises and lights flashing and banging and everything.” She had “not really wanted to carry on” and had wanted to “end everything so that nobody would have any problems.” She was slightly disinhibited, making personal comments on the examiner’s legs. The interview was considered severely impaired in quality, but ratings were made of denial of abnormality, self-neglect, nonsocial speech, drowsiness, suspiciousness, incongruous laughter, and loss of reserve.

A second interview was held the following day. She was more open this time and admitted that she thought she was being “bugged” at the first interview. She described an extensive delusional system, believing that she was being watched, was going to be killed, and was a link in an international conspiracy. She had had visual misinterpretations, and gustatory and olfactory hallucinations. She had been extremely frightened and had acted on her delusional ideas. She had abused alcohol and had made an impulsive suicide attempt. Ratings of impaired concentration and errors on cognitive testing were made.

Her husband was interviewed via a French interpreter. He said she had been depressed for some weeks before the onset of the acute disturbance, with anorexia and loss of weight. He described her overtalkativeness and her fear amounting to panic during the acute disturbance. He corroborated the account of her delusions but emphasized that they were largely consistent with the reality of life; for example, she had witnessed the sabotage of a car shortly before the onset of her illness.

In the case folder was a narrative description written by a student nurse. She described the evolution of the delusional system, including the fact that the patient accused her husband of betraying her to the authorities. When she arrived in Britain she remained awake all night talking to herself and was physically aggressive to her family. She said she was a British secret agent. On admission she smiled secretively, seemed not to hear questions addressed to her, and replied monosyllabically. When introduced to the patients in her dormitory, she was immediately drawn to one of the same age and social standing and clung to her as a child to its mother. She wore a perplexed expression and allowed herself to be led by the hand, constantly asking simple questions like “Will you take me to the toilet?” She drew simple childish sketches of her house and family, at a 4-year-old level, annotating them with words like “unhappening” and “unreligious.” In contrast to her normal personality, she was constantly going to the chapel, and wandered about in a scanty nightgown, heedless of the stares of the male patients. She behaved seductively. Her mood varied between fear, perplexity, and excitement when she would talk very quickly and giggle. All this behavior subsided completely within 2–3 days. She lost her delusions and said, “How can I have been so silly as to believe that my husband was against me?” After this rapid recovery she appeared to be a well-
adjusted, mature, and sensible woman.

None of the sources gave a full picture of this complex delusional psychosis (bouffée délirante), with its manic and paranoid elements. They each make a useful contribution. Although the student was untrained in psychiatry, her description is so clear and authentic that one has no hesitation in placing her evidence on the same level as the other data.

Ratings and the Need for Narrative. For statistical purposes, observations must be converted into numerical ratings at some stage. Commonly this is done on the spot, sometimes by junior investigators working alone. They may not record the reasons for their decisions, so that it is impossible to compare the basis for their judgments with other data. This practice stems from the belief that “instruments” like structured mental state examinations have an intrinsically high reliability. Too much emphasis has been attached to the higher interrater reliabilities—for example, coefficients of 0.73–0.77 for various editions of the Present State Examination (Wing et al. 1967; Kendell et al. 1968; World Health Organization 1973) obtained by senior co-trained researchers who knew that their reliability was being measured. This does not guarantee the reliability of ratings made under completely different conditions. It has been shown that reliability drops sharply as soon as observers believe they are no longer being monitored (Reid 1970; Taplin and Reid 1973). For this reason, it seems essential, at least in research on small samples, for all ratings to be made by two persons.

The practice of rating all information at source is also unsatisfactory because it results in too many ratings. As more and more rating scales are introduced, researchers come under pressure to complete many schedules in an effort to make their findings comparable with those of other centers. The same effect results from the use of multiple information sources. The number of ratings greatly exceeds the number of patients, and spurious correlations can readily be generated. It becomes necessary to condense the data by some form of algorithm (e.g., principal component analysis) which is surprisingly time-consuming. The combination of sophisticated mathematics with suspect data seems unwarranted.

An alternative method is to postpone ratings until the whole data base has been assembled. According to this policy, rating should be done only once, by a team of several raters, including the most senior investigators. The only material worthy of this effort is the whole body of information obtained from all sources. The raters review the patient’s dossier at one sitting, working independently at first, knowing that they may have to defend their decisions to their colleagues. Master ratings (or consensus ratings) are obtained by computation or (with more expenditure of time) by resolving disagreements. The time taken (e.g., 40 minutes per rater per dossier, and more for discussion) is considerable, but the ratings which emerge are of superior quality. This method requires the use of narrative as the medium through which evidence is compared. The interview is used not as a source of ratings but as a source of information, and it is necessary to write down what is said at the rate of about 15 words/minute while conducting it. The descriptive powers of everyone in contact with the patient are at a premium, and (unless the data can be transcribed) legible handwriting is an asset.

The use of multiple information sources does not eliminate the problem of unreliability. It may even accentuate it. If there are many sources, there are more opportunities for conflict between reports, and the rater will have to exercise his judgment when deciding which to believe. Empirically it was found in the Manchester study that the interrater reliability of ratings made from a complete dossier was not high (Downing, Francis, and Brockington 1980). It reached a satisfactory level (Pearson’s r at least .7, or Cohen’s Kappa at least .5) in only 32 of 57 items and scales. For this reason it is essential to use at least two raters when making ratings in this way.

One can debate the relative merits of number and narrative as the proper mode for recording primary observations in this field of science. The advantages of narrative are, first, that critical faculties can be used to weigh up the evidence, comparing one statement with another. Without a written record one can only average the numbers, or accept judgments on authority. Secondly, language is far richer and more flexible than any system of ratings could become, and it allows the possibility of innovation, discovering new concepts of psychopathology or new diagnostic systems. Numerical ratings are already carved in stone. Narrative, especially if it avoids professional
jargon and keeps close to events and statements, can be reinterpreted in terms of different systems of ideas. For example, the U.S./U.K. Diagnostic Project was concerned with Kraepelinian diagnosis, emphasizing characteristic symptoms like those of Schneider, and there was no intention to study other diagnostic ideas. Fortunately there was sufficient description in the schedules for Perris to make diagnoses of cycloid psychosis based on an entirely different set of pathognomonic symptoms, so that the data were used for a nosological study which was not envisaged in the original plan (Brockington et al. 1982). Once an episode has been documented in narrative form, it can be employed for diagnostic work anywhere, at any time, so long as that language is understood.

References


Acknowledgment

This research was supported in part by USPHS MH-30,938 and by the Department of Mental Health, State of Illinois. Herbert Y. Meltzer is recipient of USPHS RCRA MH-47,808.

The Authors

Ian Brockington, M.D., is in the Department of Psychiatry, University of South Manchester, Manchester, United Kingdom. Herbert Y. Meltzer, M.D., is Professor of Psychiatry at the University of Chicago and Director of the Laboratory of Biological Psychiatry, Illinois State Psychiatric Institute, Chicago, IL.