Brit. J. Psychiat. (1981), **138**, 273–282

Psychiatric Interviewing Techniques: I. Methods and Measures

M. RUTTER and A. COX

Summary: An account is given of the overall strategy and measures used in a three-phase study of styles and techniques employed in the initial diagnostic interviews with the parents of children referred to a child psychiatric clinic. The measures of interview style included interviewer activity and talkativeness, directiveness, types of questions and statements, interventions designed to elicit or to respond to feelings, and non-verbal qualities. The informant's response and the interview 'outcome' were assessed through measures of the quantity and quality of factual information obtained, and of the extent of expression of emotional feelings by the informant. Good inter-rater reliability was achieved with most measures. Some difficulties were experienced in achieving comparable thresholds for the recognition of expressed emotions.

The clinical interview constitutes the main tool for both assessment and treatment in psychiatry. As a consequence it might be thought that there would have been a great deal of research to determine the most effective ways of interviewing patients and their families for different purposes. In fact, although many practitioners have advocated a variety of approaches and methods on the basis of their personal experience and preferences (for example, Kanner, 1957; Mayer-Gross *et al*, 1969; Gill *et al*, 1954; Menninger, 1962; Sullivan, 1954; Bessell, 1971; Feusterheim, 1972), there has been surprisingly little systematic study of psychiatric interviewing techniques for clinical assessment purposes (Cox and Rutter, 1977). However, some relevant data are provided by six different types of research.

First, there have been many studies of research interviewing in *non*-clinical contexts such as survey work (see reviews by Richardson *et al*, 1965; Cannell and Kahn, 1968; Cox and Rutter, 1977). These investigations have provided many useful findings on the effects of different interviewer techniques, but there are major difficulties in generalizing the results to clinical situations. Interviewing volunteers about what they have just seen in a film or trying to get psychology students to talk about their sexual behaviour is just not the same as interviewing families who have come to a clinic to seek help about their emotional problems.

Second, there have been a variety of studies examining personal interaction in social settings (Argyle, 1972), the expression of emotions and feelings (Ekman *et al*, 1972), the development and teaching of social skills (Trower *et al*, 1978), verbal conditioning (Gelder, 1968; Williams, 1964), dyadic communication (Siegman and Pope, 1972; Beattie, 1978; Rutter *et al*, 1978), and imitation (Flanders, 1968). All of these different topics have an obvious relevance for the study of clinical interviewing skills but, again, direct clinical extension of the findings is unwarranted.

Third, there have been a variety of studies of psychotherapeutic interviewing and counselling. The work of Truax and his colleagues (Truax and Carkhuff, 1967) has been most informative in indicating the importance of therapist qualities such as warmth, empathy, genuineness and understanding (see Mitchell *et al*, 1977 for a recent review of research in this area). In this connection it is pertinent to note that these qualities are *not* unalterable facets of personality, since they may be influenced by training. Other investigators have considered features such as verbal activity and directiveness (Heilbron, 1971; Lennard and Bernstein, 1960; Shapiro and Budman, 1973). In most cases the effects have been examined in terms of therapeutic response or attendance for treatment and little is known about their merits and demerits in terms of eliciting the information and feelings needed for diagnostic appraisal.

Fourth, our own earlier studies on psychiatric and family interviewing for research purposes emphasized that the initial interview serves several quite different

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

purposes: (a) it is a means of asking questions to obtain factual information on historical events, happenings and activities; (b) it serves as a stimulus to elicit emotions, feelings and attitudes; and (c) it begins to establish a relationship which will constitute the basis for further therapeutic contact. A variety of interview techniques and strategies were developed for these purposes and it was shown that they could be combined readily. It was found possible to train both psychiatrists and non-medical workers to interview in ways which produced comparable measures of good reliability and validity on features as disparate as psychiatric symptomatology (Graham and Rutter, 1968; Rutter and Graham, 1968; Rutter *et al*, 1975), family relationships (Quinton *et al*, 1976), and feelings and attitudes (Brown and Rutter, 1966; Rutter and Brown, 1966). Moreover, the methods have proved to be utilizable by other workers (Vaughn and Leff, 1976). We were able to show that the interview is a most versatile and useful tool which can be used for purposes as different as assessing psychiatric state or measuring emotions and attitudes as subtle as warmth and hostility. However, none of the studies were designed to determine *which* interview techniques were most effective for *which* purpose.

The fifth body of research concerns the various studies of structured psychiatric interviews with adult patients. Spitzer *et al* (1964; 1970) have advocated a series of standard questions followed by non-specific probes but no cross-examination. Wing *et al* (1967; 1977) on the other hand, have utilized a standard interview in which there is both systematic coverage of all relevant areas and also flexible probing and systematic cross-questioning to get an accurate picture of actual behaviour. The approach has been shown to be reliable and valid and has been successfully employed by investigators in many different parts of the world (WHO, 1973). Saghir (1971) has shown that this highly structured approach produces a more comprehensive coverage of psychiatric symptomatology than the traditional free style interview. However, with this exception there has been little systematic comparison with other styles of interviewing and no study of its power as a means of eliciting emotions.

Finally, there are the studies by Maguire and Rutter (1976), Rutter and Maguire (1976) of psychiatric interviewing by medical students. They showed that most students exhibited quite gross deficiencies in their interviewing techniques and that the most experienced students were just as bad as the novices (i.e. the standard medical training had not helped). They then went on to develop their own training methods which they found led to considerable improvements. Their work, together with the most recent research on the same topic (Brown and O'Shea, 1980), is valuable in showing the importance of teaching medical students basic interviewing skills, but the investigations were not designed to assess the merits and demerits of different interview techniques and strategies, which was the main purpose of the set of studies reported in this series of papers. The present project was planned to develop ways of measuring clinical interview styles and interviewer behaviour; to determine the effects of different interview approaches; and to establish how the various techniques could best be combined to achieve the several aims of the initial clinical diagnostic interview with the parents of children referred to a psychiatric clinic.

This paper describes the research strategy and the development and testing of measures. Two papers report the findings of a 'naturalistic' study of interviewing by psychiatric registrars as part of routine clinical work (Cox *et al*, 1981, II; Hopkinson *et al*, 1981); and the last three papers in the series describe an experimental study designed to compare the effects of four contrasting interview styles (Rutter *et al*, 1981; Cox *et al*, 1981, V and VI).

Overall Strategy

The aim was to investigate the skills and methods needed in a clinician's first interview with the parents of a child newly referred to a psychiatric clinic. Such an interview has to fulfil many different objectives, e.g. (a) an account of the child's emotions and behaviour in order to determine whether there is a psychiatric disorder (and if there is, its nature); (b) an evaluation of the qualities and characteristics of family life and relationships; (c) an appraisal of the parents' mental and physical health; and (d) an assessment of parental feelings and attitudes with respect to both the child and the referral (Rutter, 1975; Cox and Rutter, 1977).

Children referred to the Maudsley Hospital provided the sample for study. The children's department is a general child psychiatric clinic which sees children who are predominantly of school age but including a proportion of under-fives and those over 16. The children and their families are referred by general practitioners, educational psychologists and schools, with some referrals from medical specialists. The problems presented are mainly disorders of conduct and emotions, which are in some instances associated with physical disorder, mental subnormality and educational difficulties. However, the full range of child psychiatric problems is seen. This clinic was selected because its policy ensured a fairly high degree of uniformity in the objectives of the initial interview and in the procedures to be followed. The usual practice, at the time of the study, was for one or

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

both parents to be interviewed by a psychiatric registrar who was expected to follow the guidelines provided by a memorandum on history-taking. The child, of course, was interviewed as well but that part of the initial clinical evaluation was not included in the study. A typewritten summary of interview findings was prepared for each family seen and there was a standard protocol of areas to be covered by the interview and by the summary. As a result there was an unusually high degree of consistency in interview coverage of topics, so allowing an analysis of the effects of differences in the details of interview style and techniques. All registrars had a minimum of eighteen months psychiatric training and clinical experience in the Bethlem–Maudsley Hospital training scheme.

A three phase research strategy was adopted. In the first phase, appropriate measures were developed and their inter-rater reliability was assessed. The literature was searched for methods and techniques which might be suitable for the study purposes. Discussions were held with experienced interviewers on the features of interview style thought to be crucial and many clinical interviews were observed. On this basis a systematic range of measures of interviewer behaviours and informant responses were developed and tested. Inter-rater reliability was then assessed through the analysis of 12 initial diagnostic interviews.

The second phase consisted of a naturalistic study of 36 interviews undertaken by 7 different psychiatric registrars. The registrars were asked to interview in their usual fashion and there was no attempt to get them to follow any particular strategies, styles or techniques. This part of the study was used to determine the range of approaches ordinarily followed in routine clinical practice; to examine the extent to which individual psychiatrists were consistent over different interviews in using similar styles with each informant; and to look for associations between these styles and the informant's response as reflected in the factual information given and the feelings and attitudes shown.

The third phase involved an experimental study in which interview styles were systematically varied. The findings from the naturalistic study were used to determine which interview techniques and strategies seemed most effective in eliciting emotions and factual material. These interview methods were then utilized to derive four sharply contrasting styles which reflected techniques of supposedly differing effectiveness according to the naturalistic study findings and yet which were close to styles recommended by influential teachers and practitioners in the psychiatric field. The four experimental styles were employed by two experienced interviewers who were trained in their use to a point when systematic measures showed the styles to be demonstrably different in the required ways.

Mothers attending the Maudsley Hospital with newly referred children were each interviewed twice (with a two-week gap between interviews), once by each interviewer using a different style. A balanced design ensured that each interviewer did an equal number of first and second interviews and an equal number of each of the four styles. The styles were paired in such a way (Rutter *et al*, 1981) as to allow controlled comparison of the effects of each of the four styles.

The interviews in all three stages were undertaken in the same video recording studio within the Maudsley Hospital Children's out-patient Department. The room was informally arranged with comfortable chairs in comparable fashion to other interview rooms in the department. The arrangements were always discussed in detail with families before the interview and written permission was obtained for video and audio tape recording. However, the recording equipment was very unobtrusive and most informants rapidly relaxed and soon behaved as if they had forgotten about the recording. Two linked cameras were used; one was near the ceiling on top of a high cupboard and the other was behind a one way screen. Both were operated from an adjacent room so that the camera operator was out of sight. A split screen image was used throughout with the interviewer filmed from one camera and the informant(s) from the other. The entire session was recorded using both video and audio tape. The audio tape was transcribed by a secretary trained to type all vocal and verbal material including 'guggles' and attempted interruptions. Counts and ratings were made from either the typescript or by reviewing the video tape or by a combination of the two (see below for details). All tapes were kept under lock and key and were treated with the same confidentiality as patients' case notes and records.

Development and testing of measures

Following extensive pre-testing and piloting of a wider range of scales, a comprehensive battery of measures was established with operational definitions for each measure. Inter-rater reliability was examined by two workers trained in the techniques used (A. Cox and K. Hopkinson) independently rating 12 video and audio-tape recorded interviews with the parents of newly referred child patients. Cases were taken sequentially as they were booked into the clinic; only one parent refused to be video taped. Because it was necessary not to interfere with the usual clinic procedures, the interviews were sometimes with one parent and sometimes with both parents together. However,

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

cases in which the whole family including children were seen together were excluded, as were parents who were immigrant or who did not speak colloquial English at home.

Pilot studies showed that in most cases the interview styles remained consistent throughout the interview. Because of this, because interviews varied in length from 35 to 90 minutes, and because it took an extremely long time to provide scores and ratings on all measures for the entire interview, it was decided to use only the first 30 minutes of each interview for most measures when assessing reliability.* Six different psychiatric registrars were involved in the 12 interviews so that it was possible to assess reliability over a reasonably wide range of styles.

Overall levels of reliability were assessed by the use of product-moment correlations or percentage agreement measures. In addition, inter-rater differences on mean scores were examined using the paired 't' test, in order to test for possible variations in raters' threshold or sensitivity on measures requiring some judgement.

Measures of interview style

Interview styles and interviewer techniques were assessed through five main groups of measures: talkativeness, directiveness, types of questions and statements, interventions designed to elicit or respond to feelings, and interviewers' non-verbal behaviour.

(a) *Interviewer activity and talkativeness*

First, there was a series of measures of the degree of interviewer activity and talkativeness; the same measures were also applied to informants. These included (i) the number of words spoken per unit time; (ii) the number of separate floor-holdings (i.e. pieces of talk preceded and succeeded by someone else's talk); and (iii) the mean number of words per floor-holding. These measures were direct mechanical computations from the typescript and hence there was no need to test inter-rater reliability.

(b) *Directiveness*

The extent to which an interviewer actively attempted to direct the content of the interview was assessed by measures (from the typescript) of (i) number of questions raising a new topic or area of enquiry (i.e. the topic was introduced by the interviewer rather than the informant); (ii) number of probes per subject area and (iii) direct requests for detailed descriptions of a behaviour, event or activity. In addition, a count was made of the number of interruptions (i.e. interventions which broke intc uncompleted floor-holdings of another speaker). All four measures were found to be highly reliable (see Table I). However, there was some initial difficulty as a result of the difference in rater threshold on the measure of interviewer probes per subject area. Checks during the naturalistic and experimental studies showed that this problem had been corrected.

(c) *Types of questions and statements*

Thirdly, counts were made from the typescript of the different types of questions and statements used by the interviewer. Thus, counts were made of different styles of questioning thought to influence informant talkativeness, such as (i) open questions (i.e. the form of the question expects an extended reply e.g. 'what does he do when he seems upset?'(; (ii) closed questions (i.e. the question either expects a yes-no reply or a single item answer e.g. 'does he wet the bed?' or 'how often does he miss school?'); and checks (interrogatory repetitions or near-repetitions of what the informant has just said).

In addition, counts were made of certain types of questions thought to carry disadvantages such as (i) double questions (i.e. two different questions included in the same enquiry e.g. 'does he steal or truant?' so that an answer of 'yes' would not indicate which he did) and (ii) multiple choice questions (which offered a number of alternatives which might not cover the full range of possibilities e.g. 'does he have tantrums once a week or twice a week?').

TABLE I

Reliability of measures of interviewer activity

Measures	Reliability	
	r	t
Interviewer interruptions	.94	2.54*
Informant interruptions	.78	1.00
Questions introducing new topic	.88	0.00
Probes per subject area	.95	4.18**
Requests for detailed descriptions	.88	0.43

n = 12.

* P < .05.

** P < .01.

* The first 30 minutes was also used for most measures in the naturalistic study. However, except when comparisons are made with the naturalistic study, the whole interview was used in the experimental study.

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

The reliability of all measures (as shown in Table II) was acceptable.

(d) *Interventions designed to elicit or to respond to feelings*

Fourthly, assessments were made of a number of interviewer interventions usually considered to be techniques designed to elicit or respond to feelings. These included three measures obtainable from the typescript above: (i) direct requests for the informant's own feelings (e.g. 'how did you feel?'); (ii) direct requests for self disclosures (i.e. requests for emotionally laden personal information—see below for definition) and (iii) interpretations which included both reflections of feelings (e.g. 'that made you feel angry') and confrontations or linkings of emotions. If the intervention drew attention to contrasts between verbal and non-verbal behaviour, or apparently inconsistent opinions or factual material this was called a confrontation. Linkings also could involve a comparison (for example of two different situations in which similar feelings were expressed—e.g. 'you also felt like that when your husband got promoted') or the bringing to attention of simultaneous contrasting emotions or ambivalent feelings, or indicating that some behaviour seems to serve as a defence against feeling (e.g. 'you went back to work so that you wouldn't keep thinking of his illness').

Two other measures were used which relied on ratings of emotion or feelings which may be evident in non-verbal and paralinguistic features and hence required the rater to view the videotape recording in conjunction with the typescript. These were: (i) expressions of sympathy (which relied on tone of voice and facial expression); and (ii) verbal responses to the informant's expressed feelings. The latter measure relied on sequential study of informant-interviewer dialogues which constitute a crucial aspect of emotional expression.

The inter-rater correlations for these measures (see Table III) were all above 0.7 but there were significant differences in the rater's threshold for picking up non-verbally expressed emotions from a viewing of the videotape. This difficulty could not be entirely removed by further discussion and training; consequently for the main studies *both* raters made video ratings and where differences were found a consensus agreement was reached using strict applications of the operational criteria.

(e) *Interviewers' non-verbal behaviour*

The fourth method of assessing interview style consisted of a set of measures of the interviewers' non-verbal behaviour, all derived from a viewing of the videotape recording. There were a variety of time-sampled counts during the central 5 minutes of each of the first two fifteen-minute periods of the interview of specific behaviours measured contingently according to who was talking. These

TABLE II

Reliability of types of questions

Type of question	Reliability r	Reliability t
Double questions	.88	0.72
Multiple choice questions	.77	0.40
Open questions	.98	0.32
Closed questions	.62	1.73
Checks	.78	0.55
% Open questions	.89	0.71
% Closed questions	.53	1.76
Ratio Open:closed questions	.60	2.17

n = 12.

TABLE III

Reliability of measures concerning interventions for feelings

Measures	Reliability r	Reliability t
Requests for feelings	.89	2.14
Interpretations	.78	1.24
Expressions of sympathy	.92	2.55*
Cues responded to	.84	1.68
Responses to expressed feeling	.85	2.21*

n = 12. * P < .05.

TABLE IV

Reliability of measures of interviewer non-verbal behaviour

Measures	Reliability r	Reliability t
Hand gestures while talking	.94	0.86
Listening responses	.96	2.71*
Head nods while listening	.88	2.59*
Smiles while listening	.81	1.37
Reciprocal smiles by interviewer	.79	1.59
Reciprocal smiles by informant	.90	1.53
Missed looks by interviewer	.95	0.70
Warmth	.49	0.20
Sympathy	.67	2.32*

n = 12. * P < .05.

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

included: (i) hand gestures while talking; (ii) listening responses; (iii) head nods while listening and (iv) smiles while listening. Reciprocal smiles were recorded according to whether the interviewer was responding to an informant smile or vice-versa and a count was made of the number of times the interviewer failed to pick up or respond to a look from the informant. Inter-rater reliability was high in all cases (Table I) but some measures provided threshold problems and as subsequent analysis showed that the counts did not have much association with any informant responses, little use was made of most of these measures in the main studies.

Finally, overall ratings on a 5-point scale were made of the amount of warmth and of sympathy shown by the interviewer to the informant during the first two fifteen-minute periods of the interview. These last two ratings had rather poor reliability initially (.49 and .67 respectively) but this was improved prior to the naturalistic study. However, threshold differences on sympathy remained and as a result this was rated by both workers throughout and a consensus rating was used for all comparisons.

Measures of informant response and of interview outcome

Factual information

Doctors working in the Maudsley Hospital Children's Department are given a schedule of topics about which they should obtain information. This schedule was used as the basis for the definition of the scope of child symptoms, family data and developmental data which might be assessed in the study. Several approaches were used to evaluate the range and depth of factual information obtained at any stage throughout the entire interview. First, for each pre-determined topic (based on the Departmental schedule) a rating was made as to whether any information was provided by the informant (according to a scale of 0: no data; 1: dubious or uncertain data; 2: data that behaviour probably normal or absent; and 3: data that behaviour probably abnormal or present). There was perfect agreement between raters on this scale for 71 per cent of the child symptoms, 79 per cent of the family items and 74 per cent of the developmental items (see Table V).

Secondly, there were two separate measures of the quality of factual information obtained. For each of eight major commonly occurring child symptoms, the information obtained on 12 different characteristics was rated on a 3-point scale (0: no data; 1: some data; 2: good data) according to the quality of the data. The characteristics were: onset/duration, frequency, precipitating and ameliorating factors relating to course, precipitating and ameliorating factors relating to new

TABLE V

Reliability of measures of factual information

Measures	% Agreement (perfect agreement on 4 point scale)	
Range of child symptoms	71%	
Range of family data	79%	
Range of developmental data	74%	
Quality of factual information (symptoms)	71%	
Quality of factual information (family data)	76%	
	r	t
Detailed descriptions	.80	1.23

n = 12.

incidents, social impact, parents' handling of the child's symptomatic behaviour, parents' feelings about the child's symptomatic behaviour, a detailed description of the symptomatic behaviour, two examples of the symptomatic behaviour, course, whether the symptom could be discriminated from other emotional behaviour symptoms, e.g. depression versus anxiety, and whether the symptom could be discriminated from other symptoms of similar form. Thus each symptom was given a score on a 24-point scale. There was only moderate agreement on disturbed peer relationships (r = .45) but otherwise correlations generally exceeded .75. Data on the parents' marital relationship and on the mother's mental health were similarly rated. For some purposes in the naturalistic and experimental studies it was desirable to have a cruder summary rating, the scale was subdivided into three (0: score of 0–8; 1: score of 9–16; and 2: score of 17+) to give an indication of poor, fair and good information coverage. Percentage agreement between raters on this scale was generally high (e.g. 75 per cent for mother's mental health and 85 per cent for marriage) apart from peer relationships, and the overall inter-rater agreement for both symptoms (71 per cent) and family data (76 per cent) was also acceptably high.

A further measure of the quality of factual information was provided by a score on the number of detailed descriptions of behaviour provided where inter-rater agreement was also high (r = .80).

The measures considered so far all refer to predetermined topics on which the interviewers were expected to gather information. However, it was necessary also to consider clinically significant items which are idiosyncratic to the family and thus not likely to be obtained by standardized questioning.

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

For the experimental study a special scale was constructed to pick up such items. Each typescript was perused for such items and a score of '1' was given for each item outside the list of prescribed topics on which there was any clear information. The scale covered such things as acute stresses (e.g. bereavement), significant individuals (e.g. that one parent had a lover), significant events in the past (e.g. that one parent had been reared in an institution) or significant links (e.g. that when she was the same age the mother had suffered problems similar to those of the child).

Inter-rater reliability for this measure (which was added after the reliability study) was not determined as consensus ratings were used throughout.

Emotions

Emotional material was assessed in terms of the quantity and range of emotional expression elicited by the interview (rather than in terms of a detailed evaluation of family dynamics). Feelings were rated from the first 30 minutes* of the videotape and were scored only if the informant actively expressed feelings as shown by the tone of voice used and non-verbal cues such as facial expression and gesture. The words used were taken into account in making a judgement but the content of what was said was not considered sufficient to rate feelings in the absence of non-verbal cues. Feelings were scored separately according to whether the tone was predominantly positive or negative. As many analyses were to be sequential in type, reliabilities were separately calculated in terms of the number of particular interviewer interventions followed by the informant's expression of feelings.

Inter-rater reliability was generally good but there were significant differences in threshold on several of the measures (Table VI). Because of this, consensus ratings were used in the main study.

Finally, one factual measure, the number of self-disclosures, dealt with emotionally laden personal information. These related to the mental health of immediate family members, sexual behaviour either of the parents or involving the child, violence towards the child, criminal behaviour or police contact with any member of the immediate family, tension or discord between parents or between parents and grandparents, or parental self-criticism or guilt. Once a parent had disclosed something in one of these areas there had to be a substantial increase in what was disclosed to count as further self-disclosure in that area (e.g. after maternal depression had been disclosed an account of lying awake worrying would not score but a suicidal attempt or hospital admission would count). Inter-rater reliability was high (r = 0.78) but there was again a problem with threshold differences, so that consensus ratings were used.

* The whole interview was used but prorated to a 60 minute base in the experimental study.

TABLE VI

Reliability of measures of emotions

Measures	Reliability	
	r	t
Expressed feeling per 30 mins	.89	3.57**
Negative feelings „ „ ,	.93	2.71*
Open questions followed by feelings	.84	2.71*
Interpretations followed by feelings	.76	0.76
Self-disclosures	.78	2.42*

n = 12.
* P < .05.
** P < .01.

Informants' perception of the interview

The last measure was a questionnaire completed by informants immediately after the interview. Parents were asked to indicate on a multiple choice five-point scale such perceptions as whether they felt that the interviewer was interested in them; whether the interviewer understood their problem; whether they felt relaxed or tense; whether their worries about the problem had been reduced or increased by the interview; and whether they found the interview helpful.

Discussion and Conclusions

Many of the earlier studies of clinical interviewing skills have relied on general ratings regarding such qualities as interviewer 'rapport' or 'organization' when comparing different interviews (Brown and O'Shea, 1980). Reasonable inter-rater agreement is possible on these overall ratings because experienced clinicians sharing a similar approach to interviewing show some consistency in their appraisal of what are 'good' and 'bad' interviews. However, such an approach, although valuable in assessing novices' responses to training on interviewing has two major drawbacks. First, because the ratings tend to be based on a combination of what the interviewer does and how the informant responds, it is difficult to relate interview styles to informant responses in a non-tautological fashion. Second, because the ratings of the interview are based on quality rather than style, it is not possible to determine which interview techniques

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

are most effective for which purposes. In order to examine that issue, it is necessary to have measures of different types of interview style and different kinds of interview techniques quite separate from measures of informant emotional response and of the quantity and quality of factual information obtained. The development and testing of such a range of measures was the objective of this first phase of the three-phase investigation of clinical diagnostic interviewing.

It proved possible to develop an adequate range of measures of interview style and of interview techniques. Inter-rater reliability was generally good and the measures provided a good differentiation between different interviews. There was some initial difficulty with differences in rater threshold on counts of probes per subject area but this proved to be the result of an insufficiently explicit definition. Once this had been provided, there was then no threshold difference between raters on this measure.

Somewhat greater difficulty was experienced in obtaining comparable rater thresholds for some of the measures of interviewer non-verbal behaviour and of responses to the informant's expression of feelings. It was not usually that the inter-rater correlation was low; to the contrary most exceeded 0.80. Rather it was a problem in calibration. Raters readily agreed on the criteria by which, for example, expressions of sympathy should be recognized. Raters used the same cues and, on the whole, tended to use them in the same way. The problem lay in deciding how much sympathy had to be present for sympathy to be coded. Essentially, of course, this type of variable is on a continuum and when measured on a continuous scale, acceptable levels of inter-rater reliability can be obtained (Brown and Rutter, 1966; Rutter and Brown, 1966). However, such reliable ratings are usually based on measures which reflect quite extended portions of interview rather than on categorical measures of the qualities present in a single utterance or single floor-holding, as was necessary in the present study.

Similar issues arose in the measurement of factual information provided by the informant and of the feelings expressed. No problems were evident with respect to the former and a satisfactorily high level of reliability was readily obtained. However, there were threshold difficulties in the measurement of emotions. The repeated finding of a high inter-rater correlation coefficient combined with significant inter-rater differences on threshold emphasizes the importance of always separately examining these two rather different aspects of reliability.

Most research on the recognition of emotions has used rather artificial material and settings (Argyle, 1972). Nevertheless, it appears that there is considerable agreement between people in the interpretation of emotional cues (Ekman *et al*, 1972), although individuals differ in their ability to decode non-verbal emotional information (Davitz, 1964). Clearly, direct observation of the interviews as they occur is likely to provide the fullest information on emotional expression, but we found that interactions moved too quickly to obtain adequate reliability from this method.

We therefore changed to a reliance on the analysis of videotape recordings. Something is lost in the use of black and white film, for example skin colour is absent and it is very difficult to perceive moistening of the eyes or perspiration. Nevertheless, these losses are more than offset by the gains obtained from the facility to re-run segments of interaction which are initially difficult to rate. Other difficulties in rating emotions stemmed from the expression of more than one emotion at a time (thus personal distress might accompany irritation with the child), and from the need to rate more than one informant in the joint interviews with both parents together. In general, inter-rater reliability was better with single parent interviews.

Despite these various problems, it proved possible to develop good measures of interview style and technique and informant response which were both reliable and discriminating. With those measures which showed continuing differences between raters in the threshold of recognition of emotions, consensus ratings were employed.

Acknowledgements

The research reported in this series of papers was supported by a grant from the Social Science Research Council. The papers were prepared while Michael Rutter was a Fellow at the Center for Advanced Study in the Behavioral Sciences: financial support was provided by the William T. Grant Foundation, The Foundation for Child Development, The Spencer Foundation and The National Science Foundation (BNS 78-24671).

The study would not have been possible without the active collaboration and support of members of the Children's Department of the Maudsley Hospital. Particular thanks are due to those registrars who allowed themselves to be filmed during the first two stages of the study.

Ms K. Hopkinson took a key role in the early planning and development of the first two phases of the research as well as the scoring and data analysis. She contributed importantly to the preparation of this first paper.

References

ARGYLE, M. (1972) *The Psychology of Interpersonal Behaviour*, (2nd Edition). Harmondsworth: Penguin.

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

BEATTIE, G. W. (1978) Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology*, **17**, 7–16.

BESSELL, R. (1971) *Interviewing and Counselling*. London: Batsford.

BROWN, J. E. & O'SHEA, J. S. (1980) Improving medical student interviewing skills. *Paediatrics*, **65**, 575–8.

BROWN, G. W. & RUTTER, M. (1966) The measurement of family activities and relationships: a methodological study. *Human Relations*, **19**, 241–63.

CANNELL, C. F. & KAHN, R. L. (1968) Interviewing. In *Handbook of Social Psychology*, (eds. G. Lindzey and F. Aronson), 2nd Edition, Vol 2. Cambridge, Massachusetts: Addison-Wesley.

COX, A., HOLBROOK, D. & RUTTER, M. (1981) Psychiatric interviewing techniques: VI. Experimental study; eliciting feelings. *British Journal of Psychiatry*, in press.

—— HOPKINSON, K. & RUTTER, M. (1981) Psychiatric interviewing techniques: II. Naturalistic study; eliciting factual information. *British Journal of Psychiatry*, **138**, 283–91.

—— & RUTTER, M. (1977) Diagnostic appraisal and interviewing. In *Child Psychiatry: Modern Approaches*, (eds. M. Rutter and L. Hersov). Oxford: Blackwell Scientific.

—— —— & HOLBROOK, D. (1981) Psychiatric interviewing techniques: V. Experimental study; eliciting factual information. *British Journal of Psychiatry*, in press.

DAVITZ, J. R. (1964) *The Communication of Emotional Meaning*. New York: McGraw-Hill.

EKMAN, P., FRIESEN, W. V. & ELLSWORTH, P. (1972) *Emotion in the Human Face. Guidelines for Research and an Integration of Findings*. Oxford: Pergamon.

FEUSTERHEIM, H. (1972) The Initial Interview. In *Clinical Behaviour Therapy*, (ed. A. A. Lazarus). London: Butterworths.

FLANDERS, J. P. (1968) A review of research on imitative behaviour. *Psychological Bulletin*, **69**, 316–37.

GELDER, M. G. (1968) Verbal conditioning, as a measure of interpersonal influence in psychiatric interviews. *British Journal of Social and Clinical Psychology*, **7**, 194–209.

GILL, M., NEWMAN, R. & REDLICH, F. C. (1954) *The Initial Interview in Psychiatric Practice*. New York: International Universities Press.

GRAHAM, P. & RUTTER, M. (1968) The reliability and validity of the psychiatric assessment of the child. II. Interview with the parent. *British Journal of Psychiatry*, **114**, 581–92.

HEILBRON, A. B. (1971) Female preference for therapist initial interview style as a function of 'client' and therapist social role variables. *Journal of Counseling Psychology*, **18**, 285–91.

HOPKINSON, K., COX, A. & RUTTER, M. (1981) Psychiatric interviewing techniques: III. Naturalistic study; eliciting feelings. *British Journal of Psychiatry*, in press.

KANNER, L. (1957) *Child Psychiatry*, (3rd Edition). Springfield, Illinois: Charles C. Thomas.

LENNARD, J. L. & BERNSTEIN, A. (1960) *The Anatomy of Psychotherapy*. New York: Columbia University Press.

MAGUIRE, G. P. & RUTTER, D. R. (1976) History-taking for medical students. I. Deficiencies in performance. *Lancet*, *ii*, 556–8.

MAYER-GROSS, W., SLATER, E. & ROTH, M. (1969) *Clinical Psychiatry*, (3rd Edition). London: Baillière, Tindall and Cassell.

MENNINGER, K. A. (1962) *A Manual for Psychiatric Case Study*. New York: Grune and Stratton.

MITCHELL, K. M., BOZARTH, J. D. & KRAUFT, C. C. (1977) A Reappraisal of the Therapeutic Effectiveness of Accurate Empathy, Non-Possessive Warmth and Genuineness. In *Effective Psychotherapy: A Handbook of Research*, (eds. A. S. Gurman and A. M. Razin). Oxford: Pergamon.

QUINTON, D., RUTTER, M. & ROWLANDS, O. (1976) An evaluation of an interview assessment of marriage *Psychological Medicine*, **6**, 577–86

RICHARDSON, S. A., DOHRENWEND, B. & KLEIN, D. (1965) *Interviewing*, New York: Basic Books.

RUTTER, D. R. & MAGUIRE, G. P. (1976) History-taking for medical students. II. Evaluation of a training programme. *Lancet*, *ii*, 558–60.

—— STEPHENSON, G. M., AYLING, K. & WHITE, P. A. (1978) The timing of looks in dyadic conversation. *British Journal of Social and Clinical Psychology*, **17**, 17–22.

RUTTER, M. (1975) *Helping Troubled Children*. Harmondsworth: Penguin.

—— & BROWN, G. W. (1966) The reliability and validity of measures of family life and relationships in families containing a psychiatric patient. *Social Psychiatry*, **1**, 38–53.

—— COX, A., EGERT, S., HOLBROOK, D. & EVERITT, B. (1981) Psychiatric interviewing techniques. IV. Experimental study: four contrasting styles. *British Journal of Psychiatry*, in press.

—— —— TUPLING, C., BERGER, M. & YULE, W. (1975) Attainment and adjustment in two geographical areas. I. The prevalence of psychiatric disorder. *British Journal of Psychiatry*, **126**, 493–509.

—— & GRAHAM, P. (1968) The reliability and validity of the psychiatric assessment of the child. I. Interview with the child. *British Journal of Psychiatry*, **114**, 563–79.

SAGHIR, M. T. (1971) A comparison of some aspects of structured and unstructured psychiatric interviews. *American Journal of Psychiatry*, **128**, 180–4.

SHAPIRO, R. J. & BUDMAN, S. H. (1973) Defection, termination and continuation in family and individual therapy. *Family Process*, **12**, 55–67.

SIEGMAN, A. W. & POPE, B. (1972) *Studies in Dyadic Communication*. London: Pergamon.

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press

—— ENDICOTT, J., FLEISS, J. L. & COHEN, J. (1970) The psychiatric status schedule. *Archives of General Psychiatry*, **23**, 41–55.

SPITZER, R. L., FLEISS, J. L., BURDOCK, E. I. & HARDESTY, A. S. (1964) The mental status schedule: rationale, reliability and validity. *Comprehensive Psychiatry*, **5**, 384–95.

SULLIVAN, H. S. (1954) *The Psychiatric Interview*. London: Tavistock Publication.

TROWER, P., BRYANT, B., ARGYLE, M. & MARZILLIER, J. (1978) *Social Skills and Mental Health*. London: Methuen.

TRUAX, C. B. & CARKHUFF, R. R. (1967) *Towards Effective Counseling and Psychotherapy: Training and Practice*. Chicago: Aldine.

VAUGHN, C. & LEFF, J. P. (1976) The measurement of expressed emotion in the families of psychiatric patients. *British Journal of Social and Clinical Psychology*, **15**, 157–69.

WILLIAMS, J. (1964) Conditioning of verbalizations: a review. *Psychological Bulletin*, **62**, 383–93.

WING, J. K., BIRLEY, J. L. T., COOPER, J. E., GRAHAM, P. & ISAACS, A. D. (1967) Reliability of a procedure for measuring and classifying 'present psychiatric state'. *British Journal of Psychiatry*, **113**, 499–515.

—— NIXON, J. N., MANN, S. A. & LEFF, J. P. (1977) Reliability of the PSE (ninth edition) used in a population study. *Psychological Medicine*, **7**, 505–16.

WORLD HEALTH ORGANISATION (1973) *Report of the International Pilot Study of Schizophrenia, Vol* 1. Geneva: World Health Organisation.

Michael Rutter, M.D., F.R.C.P., F.R.C.Psych., *Professor of Child Psychiatry, Department of Child and Adolescent Psychiatry, Institute of Psychiatry, De Crespigny Park, Camberwell, London SE5 8AF*

Antony Cox, M.Phil., M.R.C.P., F.R.C.Psych., *Consultant Psychiatrist, The Maudsley Hospital, Denmark Hill, London SE5 8AZ*

(*Received 11 June 1980*)

https://doi.org/10.1192/bjp.138.4.273 Published online by Cambridge University Press