Journal of Family Therapy (1979) **1**: 291–312

Clinical assessment of family interaction: a reliability study

Warren Kinston, Peter Loader and Jackie Stratford*

A standardized Current Family State Assessment was developed for possible use by family therapists. This study investigated the ability of clinicians to agree on thirty commonly used clinical and research categories. A psychiatrist and a social worker received a brief period of training and then independently rated thirty-two whole family interviews, using twelve 'non-labelled' families and eleven families referred for psychiatric disturbance in the child. Twenty-five categories were rated with moderate reliability over all families, and in any one interview the raters produced very similar rating profiles. Potential sources of unreliability are discussed. The inclusion of a systematic description of whole family interaction in routine psychiatric assessment is recommended.

No systematic clinical assessment of family interaction is widely accepted or available for use by researchers in the no longer new field of family therapy (Howells, 1962; Ackerman, 1962; Boszormenyi-Nagy and Framo, 1965; Jackson, 1968; Zuk and Boszormenyi-Nagy, 1968; Beels and Ferber, 1969; Haley, 1971; Minuchin, 1974).

Psychiatric researchers have been concerned to learn about the family relationships of their patients. However, careful studies (Brown and Rutter, 1966; Rutter and Brown, 1966), aimed at establishing the individual interview as a reliable and valid instrument to measure family life and relationships revealed it to be unsatisfactory just on those issues of most interest to family therapists, e.g. decision-making, conversation, quarreling. Brown and Rutter (1966) reviewed the literature and concluded that a direct observational approach was essential.

Cromwell *et al.* (1976) in their recent review of methods of assessing

* The Academic Department of Child Psychiatry, Institute of Child Health, University of London, London, U.K., and The Department of Psychological Medicine, The Hospital for Sick Children, Great Ormond Street, London, U.K.

Requests for reprints to Dr P. J. Loader, The Department of Psychological Medicine, The Hospital for Sick Children, Great Ormond Street, London WC1N 3JH, U.K.

0163–4445/79/030291+22$02.00/0

families referred to reports by therapists of their observations of families and stated 'rarely are systematic observational data collected'. A recent project (Kinston and Bentovim, 1978) which attempted to determine specific changes in families following brief focal family therapy found itself handicapped by the absence of a standardized descriptive vocabulary. The G.A.P. Report (1970) on the field of family therapy remarked on the absence of a common vocabulary for the description of family interaction and urged its development as a priority. Researchers have developed categorizations and descriptive terms to meet their own interests (Broderick, 1971; Riskin and Faunce, 1972; Haley, 1972) but the specific problems of clinical research and routine clinical use have not been tackled.

Some system is needed to permit a reduction and summary of data, to allow comparisons between different families and, over time, within a family, and to guide the thinking and observation of newcomers to the field. This paper presents a set of categories, together with data on reliability, for rating family interactions as observed during a clinical interview.

Review of the literature

Straus (1969) has collected systematic measures of the family, most of which have limited clinical relevance; and family therapists (Glick and Kessler, 1974) have offered detailed outlines for the clinical evaluation of the family, which are useful but combine history, observation and interpretation non-systematically.

Wells and Rabiner (1973) describe a schedule (the Family Index of Tension) developed in association with an interview which was oriented towards the planning of treatment. Ratings were made of individuals and dyads in relation to the family and not of the family-as-a-whole. The data obtained by the Henry Ittleson Center Family Interaction Scales (Behrens, *et al.*, 1969), devised for use with home visiting, mostly concern dyadic interaction but do include several ratings of 'Family Group Patterns'. Riskin (1976) has made ratings using categories from his research Family Interaction Scales (Riskin and Faunce, 1970) in a series of informal interviews of two 'non-labelled' families, but no evidence was offered for reliability.

Clinical assessment

One aim of the present study was to widen the scope of clinical assessment of the child to include data on family interaction. A typical clinical assessment in the U.S.A. (Chess, 1959) and U.K. (Maudsley Hospital, unpub-

lished manual) has not involved consideration of the family as a single unit, i.e. it omitted the direct assessment of family interaction, which we will refer to as the 'family state assessment'.

Artificial stimulation of family interaction is unnecessary within a department of child psychiatry where families are routinely seen as part of the diagnostic assessment of referred children and where family therapy is a common modality of treatment. In these settings families enact, cope with and discuss conflicts which have high personal significance and the patterns of interaction observed make clinical sense in terms of historical details and presenting disturbances (Bentovim and Kinston, 1978). The family state assessment can be grouped with the physical examination and the mental state examination. In common with these, it must be (1) performed by a professional (2) who has been suitably trained and (3) accepts certain rules for categorizing data (4) as he observes and probes the object of his investigation. There are other features in common. The family state assessment is performed at a particular point in time and hence any finding may be highly context-dependent. Further, assessment of any one aspect of the state must be considered in relation to the whole picture which will include historical and other details. Finally, assessment of reliability and validity of a family state requires the development of a formal standardized procedure. Our purpose in this study was to determine whether clinicians could agree on the family interaction they were observing during typical clinical interviews.

Materials and methods

Pilot study

The research and clinical literature was scanned for categories or items of interaction and thirty-three which appeared to be suitable for clinical use were chosen. A glossary was constructed which consisted of definitions of the categories and anchor descriptions for the 1st, 3rd and 5th points of a 5-point ordinal rating scale. A preliminary pilot study was performed using many members of a department of child psychiatry as raters. This pilot included psychiatrists, psychologists, social workers, psychotherapists and students. They were asked to rate any family interview where family interaction was a feature. Many interviews were diagnostic assessments and ratings were made by those behind a one-way screen as well as by one or both of the interviewers. One hundred and thirty ratings of fifty-five families were made by twenty-eight raters. All first ratings were discarded as 'practice attempts' which left 102 ratings of forty-three families by

twenty-two raters. Fourteen of these were second and third ratings of families attending regularly. The raters were subsequently asked to comment on the definitions and descriptions in the glossary, the ease of use of the schedule and the relevance of the rating. Results of the pilot were encouraging. Almost all items showed a spread over the whole 5 scale-points and none spread less than 4. A Monte Carlo technique for analysing the data for inter-rater reliability and discrimination of families category by category was developed by D. Boniface*. The eighty-eight initial ratings of families were studied using it: twenty-seven of the thirty-three categories were being rated similarly by raters watching the same interview at a level of $P < 0.015$, and families were well-discriminated in each of twenty-eight categories. Some evidence for intra-rater reliability was seen over a three-monthly interval from families who were not changing clinically.

The rating schedule was modified to take the findings of the pilot into account. The authors then began rating family interviews of consecutively-referred families and discussing the ratings, especially discrepancies. Video-tape recordings were also used and repeated observations, ratings and discussions with colleagues produced further refinements and rules for the ratings. The third edition of the schedule, containing thirty items, resulted and this was used in the study reported here.

Design of main study

In order to obtain meaningful reliabilities it is necessary that the scales be used over all or most of their range. To ensure this, two groups of families were used. There was a total of twenty-three different families and thirty-two different interviews.

The first group, the 'Clinical Group', consisted of eleven families who were seen consecutively in routine clinical practice for diagnosis, therapy or follow-up by P.L. and J. S. as co-therapists. To be included in the study the family had to consist of at least three members with one child over two years. This group contained families with one to four children, whose ages ranged from three to seventeen. Most commonly two children attended. Two families were single parent. Although three families were not of English origin, all members spoke English fluently. Social classes I to IV were represented. Some families were seen on more than one occasion, some at two- to four-week intervals, and a total of twenty interviews were rated.

* Statistician, Institute of Child Health, University of London, London, U.K.

The second group of families, the 'Eczema Group', contained at least one child attending at a dermatology clinic with eczema. The group was collected from consecutive attendances at the Clinic. The criteria for exclusion were any member currently labelled as psychiatrically ill, the absence of a child between two and thirteen years of age, single-parent families, and inability to speak English fluently. Suitable families were invited to attend for a standard interview until twelve families had agreed. There were twenty-two refusals. The composition of this group was similar to the clinical group. P.L. and J.S. each interviewed six families and whilst one interviewed the other observed using closed-circuit television with the consent of the families.

Procedure

Immediately following the interview the two raters (P.L. and J.S.) completed the *Current Family State Assessment** (C.F.S.A.) rating schedule independently. Data was then transferred to punched cards for computer analysis. Analysis included an inter-correlation matrix to look for item specificity. During the three months of data collection, raters were not given access to previously completed ratings and were required not to discuss the use of the schedule, ratings of the families in the study, or ratings of other families at the clinic.

The C.F.S.A. rating schedule

General instructions

The rater is asked to take an outsider's point of view, to consider the family group as a whole unit and to be not too influenced by any one member. Ratings are based on observations rather than inferences or interpretations and must be made on the basis of the whole interview. Rating is performed immediately after the interview by scoring the appropriate number on the 5-point ordinal scales in the C.F.S.A. The schedule incorporates a glossary and consulting it and rating can be completed in ten to twenty minutes once familiarity is attained.

The categories

The categories chosen fall into two main groups. Group A categories (1 to 20) are derived principally from the research literature; they tend to

* Copies of the C.F.S.A. are available on request.

be relatively simply defined and directly observable, and broadly cover the areas of communication, relationship and affects. Group B categories (21 to 30) are of clinical origin; they tend to require more judgments and cover complex areas such as coalitions, boundaries and consensual experience. All categories are descriptions of behaviour in the interview rather than of enduring characteristics of the family. Such characteristics could be assessed by noting the stability or fluctuations of categories over time. Brief descriptions of the categories are provided in the Appendix.

Results

In the analyses that follow, the data have been grouped in two different ways to reduce bias. When general trends or tendencies were being checked for, all ratings in the study were included, i.e. thirty-two interviews of twenty-three families. Counting the same families more than once could produce a misleadingly inflated impression of reliability either due to greater familiarity with the family by the rater or from intervening discussions between the raters who were working together. So, for the assessment of reliability only the first rated interviews have been used, i.e. $N = 23$.

We wished to assess the extent and nature of inter-observer agreement on the thirty 5-point scales which comprised the C.F.S.A., both to check the value of this approach to family measurement and as a guide to further development of the scales. The simplest method is to count the number of disagreements of different degrees (the '*D*' scores) as in Table 1.

There are problems in interpreting reliability scores. Firstly, if the full range of the scale has not been used because the families are not different, the *D* scores do not really test the scale and high agreement must remain suspect. (The range of the scale used is evident from Table 2.) Secondly, agreement may be high in one part of the scale, or most of it, but poor in another part. Thirdly, some families or interviews might have been characterized by a very high agreement between raters, irrespective of the particular items, while other families were generally disagreed about. Fourthly, a rater may have rated slightly differently in a consistent manner, e.g. always one point higher. Finally, if there were any substantial differences in family interaction ratings between the eczema and clinic groups, an artefactually high rate of reliability could have been achieved simply by both raters making systematic differences between the groups but rating unreliably within the groups. This last possibility was checked and excluded: within-group agreement was generally similar to the agreement with the groups combined.

TABLE 1. *Extent of rater disagreement*

No.	Category	*D*: no. of points disagreement 0	1	2	3	4	% when $D = 0$ or 1
1	Clarity	10	13	0	0	0	100
2	Continuity	12	10	1	0	0	96
3	Acknowledgment	13	9	0	1	0	96
4	Information exchange	10	12	1	0	0	96
5	Interruptions	14	6	3	0	0	87
6	Laughter	12	10	1	0	0	96
7	Equality of participation	14	7	2	0	0	91
8	Self-affirmation	17	6	0	0	0	100
9	Request for commitment	14	7	2	0	0	91
10	Agreement	10	9	4	0	0	83
11	Disagreement	10	10	3	0	0	87
12	Positive support	8	9	6	0	0	74
13	Attack	10	9	4	0	0	83
14	Intrusiveness	5	15	3	0	0	87
15	Mind-reading	8	14	1	0	0	96
16	Affects—range	10	11	1	1	0	91
17	Affects—intensity	7	14	2	0	0	91
18	Tension	10	11	2	0	0	91
19	Comfort	12	10	1	0	0	96
20	Humour	12	10	1	0	0	96
21	Effectual parental coalition	11	10	2	0	0	91
22	Generational boundaries	10	9	4	0	0	83
23	Alliances	10	12	1	0	0	96
24	Resonance	12	8	3	0	0	87
25	Flexibility	12	9	2	0	0	91
26	Conflict acknowledgment	11	12	0	0	0	100
27	Feeling of safety	8	11	3	1	0	83
28	Identity struggles	13	8	2	0	0	91
29	Experience of the environment	13	10	0	0	0	100
30	Grasp of meaning	6	12	3	2	0	78

Data taken from the first rated interview of each family ($N = 23$).

To check whether raters were using the scales differently, each item was tested using the Wilcoxon Matched Pairs Signed Ranks Test (Siegel, 1956). Significant differences existed for almost half the Group A scales: *Equality of Participation* $P < 0.05$, *Agreement* $P < 0.05$, *Attack* $P < 0.01$,

Intrusiveness $P < 0.05$, *Mind-reading* $P < 0.05$, *Range of Affects* $P < 0.05$, *Intensity of Affects* $P < 0.01$, *Tension* $P < 0.01$, *Comfort* $P < 0.05$. Among Group B categories only *Grasp of Meaning* was rated differently ($P < 0.05$).

TABLE 2. *Distribution of ratings and percentages of complete agreement at each scale point*

No.	Category	Score 1 N_1	%	Score 2 N_2	%	Score 3 N_3	%	Score 4 N_4	%	Score 5 N_5	%
1	Clarity	0	—	5	40	30	40	29	48	0	—
2	Continuity	0	—	3	67	24	42	29	41	8	25
3	Acknowledgment	1	0	7	29	33	73	20	50	3	0
4	Information exchange	1	0	11	36	38	63	13	15	1	0
5	Interruptions	2	0	10	0	39	67	10	60	3	67
6	Laughter	3	0	25	32	26	54	9	67	1	0
7	Equality of participation	1	0	19	53	37	70	7	29	0	—
8	Self-affirmation	0	—	6	0	49	82	9	67	0	—
9	Request for commitment	5	40	15	53	41	78	2	0	1	0
10	Agreement	9	22	21	29	28	57	5	80	1	0
11	Disagreement	14	57	20	20	26	38	3	0	1	0
12	Positive support	5	0	18	44	28	57	10	40	3	0
13	Attack	21	57	15	27	23	52	4	0	1	0
14	Intrusiveness	6	0	18	11	27	30	7	29	6	67
15	Mind-reading	22	36	21	19	19	42	2	0	0	—
16	Affects—range	0	—	12	50	28	50	16	38	8	0
17	Affects—intensity	3	0	24	42	29	41	6	0	2	0
18	Tension	1	0	9	0	31	58	21	67	2	100
19	Comfort	2	100	22	45	25	40	12	83	3	67
20	Humour	8	50	31	71	20	40	3	0	2	100
21	Effectual parental coalition	10	40	16	50	20	50	13	46	5	0
22	Generational boundaries	9	44	13	31	25	48	13	31	4	0
23	Alliances	8	75	15	40	28	50	10	40	3	0
24	Resonance	2	0	12	50	13	46	32	63	5	0
25	Flexibility	6	33	33	61	22	45	3	0	0	—
26	Conflict acknowledgment	4	0	14	29	38	79	8	75	0	—
27	Feeling of safety	7	86	7	29	26	46	16	38	8	25
28	Identity struggles	29	62	22	36	11	36	0	—	2	100
29	Experience of the environment	2	100	3	67	33	73	19	42	7	57
30	Grasp of meaning	5	40	12	17	27	37	17	24	3	0

Data taken from all interviews of all families.
$N_1+N_2+N_3+N_4+N_5 = 64$.

From Table 1 between 74% and 100% of all ratings were either identical or within 1 point of each other. Eighteen items were used over the full 5 points and nine items over 4 points. In these categories there were no cases of 4-point disagreements and only a few isolated cases of 3-point disagreements. Three items (*Clarity*, *Self-affirmation* and *Identity Struggles*) were used over only 3 points of the scale, and D scores are more difficult to interpret. For *Intrusiveness*, *Mind-reading*, *Intensity of Affects* and *Grasp of Meaning*, there were markedly fewer complete agreements.

The complete agreements were further examined for evenness of occurrence over all scale points. Table 2 displays the frequency with which any particular score was chosen for each item in our total sample and the percentage of these which represent complete agreement. The percentage agreement varies from 0 to 100% and is mostly between 30% and 80%. There is more commonly either no agreement or a high agreement at the extremes (scores 1 or 5). The categories for which there was at least one agreement at each of the scale points used were *Clarity*, *Continuity Comfort*, *Relation to the Environment* and *Feeling of Safety*. By inspection poorer categories include those with few complete agreements mentioned above and also *Information Exchange*, *Disagreement* and *Flexibility*.

Weighted kappa (K_w) is the most stringent and suitable single statistic for the assessment of inter-rater reliability (Cohen, 1968; Hall, 1974). It is distribution-free, allows credit for partial rater agreement, corrects for rater agreement due to chance, makes use of each point of the rating scale and corrects for differences in the rater mean scores. Table 3 lists the K_w value for each item and the level of significance reached. Using $P < 0.025$ as the cut-off point, twenty-five of the thirty categories have been rated reliably. The unsatisfactory Categories are *Clarity*, *Information Exchange*, *Positive Support*, *Intensity of Affects* and *Grasp of Meaning*.

To take account of the possibility that reliability was being diminished by the global impairment in a particular interview or with a particular family, and to examine whether the raters were providing similar profiles of ratings, a Family Interview Discrepancy score (FID) was calculated by taking the average of the square of the disagreement score for each item, i.e. $\Sigma D^2/n$ (D = disagreement in points, n = number of items rated). It can be seen that FID can vary from 0 (every item scored identically by both raters) to 16 (whenever one rater rates 1, the other rates 5). Either extreme is highly unlikely. Using a Monte Carlo technique, we assessed the probability of a score of 1.55 occurring as being less than 1 in 500. Originally, we regarded FID scores over 1.00 as unsatisfactory and did not commence the study until profiles at this level were being obtained regularly.

TABLE 3. *Inter-rater reliability by weighted kappa*

No.	Category	K_w	*P* value
1	Clarity	0.19	N.S.
2	Continuity	0.41	0.001
3	Acknowledgment	0.39	0.005
4	Information exchange	0.17	N.S.
5	Interruptions	0.38	0.01
6	Laughter	0.41	0.01
7	Equality of participation	0.28	0.01
8	Self-affirmation	0.42	0.01
9	Request for commitment	0.25	0.025
10	Agreement	0.27	0.01
11	Disagreement	0.38	0.01
12	Positive support	0.15	N.S.
13	Attack	0.35	0.01
14	Intrusiveness	0.31	0.01
15	Mind-reading	0.30	0.025
16	Affects—range	0.25	0.025
17	Affects—intensity	0.13	N.S.
18	Tension	0.34	0.001
19	Comfort	0.49	0.001
20	Humour	0.42	0.001
21	Effectual parental coalition	0.54	0.001
22	Generational boundaries	0.43	0.001
23	Alliances	0.47	0.001
24	Resonance	0.34	0.01
25	Flexibility	0.30	0.025
26	Conflict acknowledgment	0.43	0.001
27	Feeling of safety	0.33	0.01
28	Identity struggles	0.34	0.025
29	Experience of the environment	0.51	0.001
30	Grasp of meaning	0.02	N.S.

Data taken from the first rated interview of each family ($N = 23$).
N.S. = Not significant.
K_w can vary from +1 (perfect agreement) through 0 (chance agreement) to −1 (complete disagreement) and thus may be interpreted like a correlation coefficient (see Text). *P* value indicates the significance of departure from 0 in the positive direction.

TABLE 4. *Family interview discrepancy score (FID): inter-rater discrepancy for all ratings of a particular family interview*

Eczema group		Clinical group	
Family	FID	Family	FID
1	0.83	1	1.17
2	0.66	2*	1.45, 1.31, 0.69
3	0.55	3	0.86
4	0.79	4*	1.14, 0.76
5	0.72	5	0.69
6	0.66	6	0.66
7	0.48	7*	1.07, 0.48
8	0.31	8	0.76
9	0.66	9*	0.72, 0.76, 0.31, 0.90
10	0.66	10*	0.62, 0.66, 1.03
11	0.79	11	0.76
12	1.55		

FID $= \Sigma D^2/n$
D = no. of points disagreement for an item
n = no. of items (29).
* These families were rated on more than one interview.

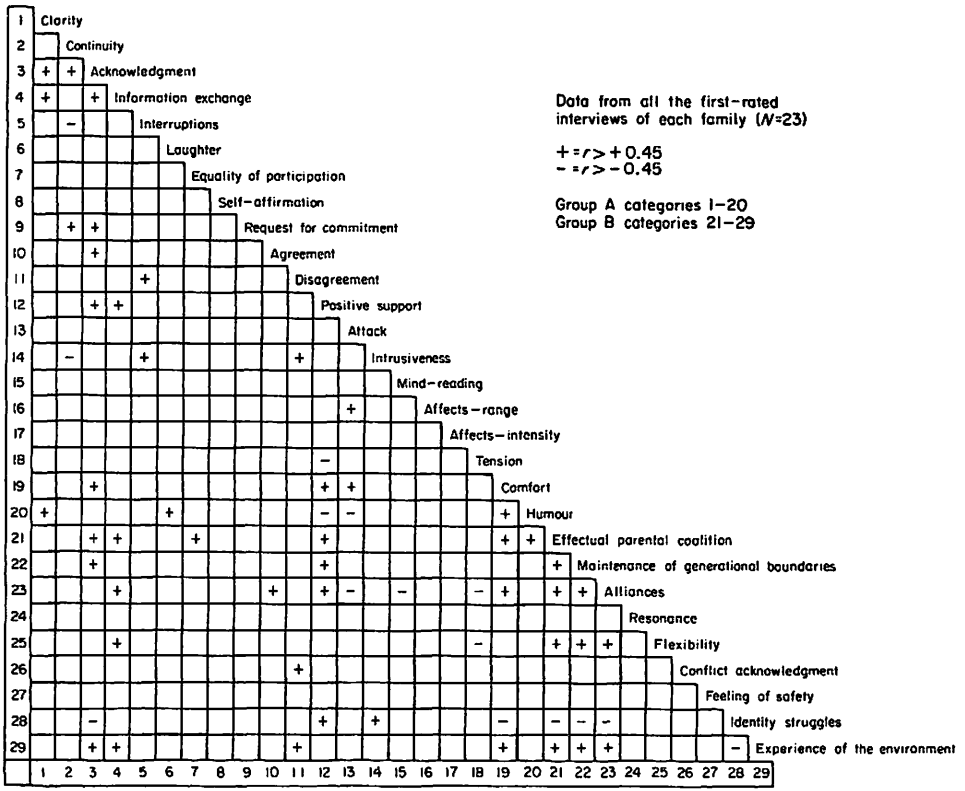
In examining FID, *Grasp of Meaning* was omitted as it was regarded as unsatisfactory and invalid by the raters. The FID scores are presented in Table 4: all scores are low. Omission of items unreliable by K_w (i.e. $n = 25$) produced slight improvements in the scores. It is therefore clear that unreliability was not due to one or a few poorly-rated interviews.

The FID varied between 0.31 and 1.55. The lowest score represents twenty-two complete agreements, six items with 1-point disagreements and one item with a 2-point disagreement. The highest score represents eight complete agreements, sixteen 1-point disagreements, five 2-point disagreements and one 3-point disagreement. As can be seen from Table 4, in those cases where there were several interviews with one family, there was a trend for rating profiles to become more similar.

Intercorrelation matrix

Scores on categories 1 to 29 were intercorrelated to check for item specificity and possible halo effects (Table 5). An arbitrary cut-off point for high correlations was taken at ± 0.45. The matrix contains 435 cells, and of

TABLE 5. *Intercorrelation matrix*



these sixty-three (14%) are highly intercorrelated. The frequency of intercorrelations within Group A (210 cells) is twenty-four (11%) and between Group A and Group B (180 cells) is twenty-six (14%) while that within the forty-five cells of Group B is thirteen (29%). The overlap within the more global clinical categories is to be expected. Three categories within Group A were particularly overlapping: *Acknowledgment*, *Positive Support*, and *Comfort*. The value of avoiding bi-polar scales is evident in the lack of high negative correlations between *Comfort* and *Tension* or *Positive Support* and *Attack*.

Discussion

Our two clinicians found rating an unusual and sometimes difficult exercise despite their active engagement in family work. Rating made observation of interviews an active and purposeful task; and it did not interfere with therapeutic understanding of the family. The coverage of the scales was not complete but the schedule was more comprehensive than informal clinical descriptions currently used in the Department. The absence of many high intercorrelations in Table 5 suggested that item specificity was satisfactory.

The raters initially felt more at home with clinical categories of Group B, but then discovered that these were no easier to rate than those derived from family interaction research (Group A). The subjective ease or difficulty of rating had little relationship to reliability but was related to bias (i.e. significant differences between the rater's mean scores). The raters were surprised to find lack of reliability on three of the unsatisfactory items (*Clarity*, *Positive Support*, *Intensity of Affects*). *Clarity* was used over a narrow range and it is difficult to draw conclusions about the scale as a whole. *Positive Support* may have been inadequately differentiated from similar categories (see Table 5).

Factors affecting reliability

(a) *Halo effect.* The intercorrelation matrix (Table 5) reveals that many categories correlate highly with others. This overlap might be a result of a halo effect or due to the conceptual and phenomenological properties of the categories. An attempt was made to minimize the halo effect in various ways.

(1) Using explicit definitions and specification of cues. For example, *Intrusiveness* has a wide variety of connotations but by restricting its

definition to particular forms of interaction it could be distinguished from *Interruptions*, *Continuity* and *Mind-reading* with which it was initially confused.

(2) Emphasis on particular observations and not general impressions. The rater was instructed to go over the interview in his mind searching for examples of the particular phenomenon and not simply to rate on his global idea of what the family was 'really' like.

(3) Developing common thresholds for rating. For example, *Range of Affects* and *Intensity of Affects* which were not reliably rated initially were re-defined on all 5 rating points. This improved the former but not the latter.

(4) Using unipolar scales, e.g. *Tension* and *Comfort*. The bi-polar scales which have been included, have been subsequently re-defined in the fourth edition as two scales each. *Resonance* is now *Enmeshment* and *Disengagement*, and *Feeling of Safety* is *Overprotection* and *Neglect*. In our experience both extremes occurred too often together to permit the use of bi-polar scales.

(5) Training raters to determine their own specific halo biases.

(b) *Difficult categories*. For many families there are a few categories which the raters experience as subjectively difficult to rate. These instances have usually heralded the further refinement and development of rules for rating. The schedule will be in a transitional form for some time while experience in describing and assessing families accumulates.

(c) *Different families*. The families on which the C.F.S.A. has been developed have a child with psychiatric problems attending at a particular institution in the U.K. Use in families containing an adult patient, a schizophrenic or autistic member, or in multi-problem families has not been attempted. The size of the family profoundly affects interaction. Only small families were studied in the reliability trial, although the piloting included families with larger numbers of children and with a single parent.

(d) *Age of the children* (*stage of family life-cycle*). This can influence reliability on a number of the measures, e.g. *Equality of Participation* can only be rated in terms of age-appropriateness. In the absence of agreed standards for interaction in different phases of the family life-cycle it is necessary to rely on the clinical experience of the rater and this was a source of error in the pilot study with less experienced clinicians.

(e) *Averaging of the interview.* A routine diagnostic interview as performed in the Department provided many of the ratings and proved satisfactory. However, when ratings were made on therapy interviews, the family often showed dramatic alterations in interaction in association with therapeutic interventions. This tended both to confuse the rater and to lead to errors in scoring because of the need to average over the whole interview.

(f) *Deviant family members.* The raters must also average the family members and this required that if one family member was markedly different from the rest of the family his contribution to the interaction had to be mentally averaged-out over the rest. The rater was instructed to consider routinely each member's or dyad's or triad's contribution before rating. Nevertheless excessive deviance did contribute to unreliable rating scores despite good agreement by raters as to what they were observing.

(g) *Family distress.* Because rating requires both careful observation and empathic understanding of the family, families which show extreme pathology can interfere with accurate rating. Raters become over-involved (whether as the interviewer or as an observer using a one-way screen or videotape) and feel the need to defend against the pain of the interview. This results in the blocking of their observation and they become influenced by halo effects and rate erratically.

(h) *Amount of emphasis placed on historical details.* Inexperienced raters tended to place excessive emphasis on what the family reported. However, to rule out all historical details unduly constricted and confused raters.

(i) *Number of raters.* An increased number of raters greatly increased the reliability of the ratings. In work ancillary to this study, a few interviews were rated by four to seven raters. Usually ratings were identical or within 1 point. When ratings spanned 3 or more points then discussion often, but not always, resolved the discrepancy and a consensus score was within 1 point. Particularly in view of the bias in Group A categories, use of several trained raters is recommended if data are to be used for research purposes.

(j) *Rater as interviewer.* We were concerned that rating while conducting interviews might be less satisfactory than when the rater was simply observing. To check this, inter-rater reliabilities were calculated separately for the clinical co-therapy interviews and the eczema interviews. Any

marked differences would then suggest this. They were not found. Also half the interviews in the Eczema Group were given by one interviewer and half by the other. There was no difference between the reliability obtained in each case and the families were described similarly by the two raters.

(k) *Intra-rater variation.* No formal intra-rater reliability testing has been performed but this can be expected to be high from our general experience. High intra-rater reliability on a formal test should not be confused with actual intra-rater reliability in the field. It was our impression that lapses in concentration, distractions and delays during rating and similar factors, could lead to gross distortions of recall and hence poor rating. The FID of 1.55 for Family No. 12 of the Eczema Group was probably due to intra-rater factors of this type.

Limitations of the results

(a) *Reliability.* This study was concerned to demonstrate that family therapists, given a brief period of training, can agree on a number of aspects of family interactional behaviour which other clinical and research workers have regarded as important in the description of families. Agreement is there but it is not impressive and not at a level of comfort for general research purposes. Although various clinicians have tried out the schedule, only three raters from one institution have formally used the method and other workers might not agree with the definitions used, the rules for ratings imposed or the ratings made. For introducing the C.F.S.A. we prepared a series of videotapes to serve as standards for rating and the new raters were given a period to rate and discuss discrepancies. They attended regular meetings to talk over problems which arise in making ratings.

(b) *Validity.* No evidence is offered at this stage for validation. Are the raters rating what the family are actually doing? Given accurate ratings, are they of any relevance to the everyday life of the family? These issues need examining. In addition we need to know whether these measures have any predictive power, or stability over time.

(c) *Diagnostic value.* The rating schedule does not allow a diagnosis to be made and does not relate to any of the primitive forms of family classification currently existing (Fisher, 1977). The restriction of ratings to observables does not result in as great a loss of information as clinicians

might initially conclude. For example, a hostile but highly defended family who are rated low on *Attack* simply because they do not attack each other at interview, might well be rated low on *Comfort*, low on *Flexibility* and low on *Clarity*, findings which clearly distinguish it from another family who also score low on *Attack* but have no 'unconscious' or covert hostility.

Conclusion

Two family therapists could agree at a clinically acceptable level on most of thirty categories of family interaction and could produce very similar rating profiles of interviews. However, our study suggests that the vocabulary of family therapy requires explicit definition. Even commonly-used terms may be being employed in idiosyncratic, excessively diffuse or biased ways. The Current Family State Assessment, though still undeveloped as a research tool, provides a more comprehensive and systematic opportunity for the clinician to consider and record family interaction than has existed heretofore. Once familiarity with the schedule is attained, the time taken to rate an interview is not excessive.

References

ACKERMAN, N. W. (1962) Family psychotherapy and psychoanalysis: the implications of difference. *Family Process*, **1**: 30–43.

BEELS, C. C. and FERBER, A. S. (1969) Family therapy: a view. *Family Process*, **8**: 280–318.

BEHRENS, M. L., MEYERS, D. I., GOLDFARB, W., GOLDFARB, N. and FIELDSTEEL, N. D. (1969) Henry Ittleson Center Family Interaction Scales. *Genetic Psychology Monograph*, **80**: 203–259.

BENTOVIM, A. and KINSTON, W. (1978) Brief focal family therapy where the child is the referred patient. 1. Clinical. *Journal of Child Psychology and Psychiatry*, **19**: 1–12.

BOSZORMENYI-NAGY, I. and FRAMO, J. (Eds) (1965) *Intensive Family Therapy*. New York. Harper & Row.

BRODERICK, C. B. (1971) Beyond the five conceptual frameworks: a decade of development in family theory. *Journal of Marriage and the Family*, **33**: 139–160.

BROWN, G. W. and RUTTER, M. (1966) The measurement of family activities and relationships. A methodological study. *Human Relations*, **19**: 241–263.

CHESS, S. (1959) *An Introduction to Child Psychiatry*. New York. Grune & Stratton.

COHEN, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**: 213–220.

CROMWELL, R. E., OLSON, D. H. and FOURNIER, D. G. (1976) Tools and techniques for diagnosis and evaluations in marital and family therapy. *Family Process*, **15**: 1–49.

FISHER, L. (1977) On the classification of families: a progress report. *Archives of General Psychiatry*, **34**: 424–433.

GLICK, I. E. and KESSLER, D. R. (1974) *Marital and Family Therapy*. New York. Grune & Stratton.

GROUP FOR THE ADVANCEMENT OF PSYCHIATRY (1970) *The Field of Family Therapy*, **7**, *Report No. 78.*

HALEY, J. (Ed.) (1971) *Changing Families. A Family Therapy Reader*. New York. Grune & Stratton.

HALEY, J. (1972) Critical overview of the present status of family interaction research. In: J. Framo (Ed.), *Family Interaction: A Dialogue between Researchers and Family Therapists*. New York. Springer.

HALL, J. N. (1974) Inter-rater reliability of ward rating scales. *British Journal of Psychiatry*, **125**: 248–255.

HOWELLS, J. G. (1962) The nuclear family as the functional unit in psychiatry. *Journal of Mental Science*, **108**: 675.

JACKSON, D. D. (1968) *Human Communication*, **1** and **2**. Palo Alto. Science and Behavior Books, Inc.

KINSTON, W. and BENTOVIM, A. (1978) Brief focal family therapy where the child is the referred patient. 2. Methodology and results. *Journal of Child Psychology and Psychiatry*, **19**: 119–143.

MINUCHIN, S. (1974) *Families and Family Therapy*. Cambridge. Harvard University Press.

RISKIN, J. (1976) 'Nonlabeled' family interaction: preliminary report on a prospective study. *Family Process*, **15**: 433–439.

RISKIN, J. and FAUNCE, E. E. (1970) Family interaction scales. 1. Theoretical framework and method. *Archives of General Psychiatry*, **22**: 504–512.

RISKIN, J. and FAUNCE, E. E. (1972) An evaluative review of family interaction research. *Family Process*, **11**: 365–455.

RUTTER, M. and 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.

SIEGEL, S. (1956) *Nonparametric Statistics: For the Behavioral Sciences*. New York. McGraw Hill.

STRAUS, M. A. (1969) *Family Measurement Techniques. Abstract of Published Instruments, 1935–1965*. Minneapolis. The University of Minnesota Press.

WELLS, C. F. and RABINER, E. L. (1973) The conjoint family diagnostic interview and the family index of tension. *Family Process*, **12**: 127–144.

ZUK, G. H. and BOSZORMENYI-NAGY, I. (Eds) (1968) *Family Therapy and Disturbed Families*. Palo Alto. Science and Behavior Books, Inc.

Appendix

Brief descriptions of C.F.S.A. Categories

(1) *Clarity*. This refers to the way communications made by family members convey meaning, but excludes aspects of articulation. Incongruence between verbal and non-verbal communication, ambiguities, contradictions, excessive subtlety, sarcasm or irony will all impair clarity. Clarity may be lost only occasionally, for example, in discussion of emotionally sensitive issues, or deficiencies in clarity may be a characteristic of the family method of communication.

(2) *Continuity.* This refers to the ability of the family both to share a focus of attention and to move smoothly from one topic to the next. The rater looks for breaking of sequences of interaction, disregard for previous statements or agreements about what is to be discussed, and the presence or absence of meaningful or conventional links between topics. Grammatical or strict logical continuity is not necessary. At the lower extreme there are many changes of topic which occur so inappropriately that the interviewer experiences confusion.

(3) *Acknowledgment.* This refers to family members providing indications to show that they have received and understood directed communications. Acknowledgment may be made either verbally or non-verbally. At the upper extreme acknowledgment is clear, automatic, directed, routine and varies from being barely noticeable to being overt as required. At the lowest level members act as if impervious to each other or disqualify directed communications.

(4) *Information exchange.* This refers to statements made by family members to each other conveying factual or historical details. It assesses how much the family members talk to each other and hence does not include information provided in direct response to a question from the interviewer, but may include discussion by the family as to the correct response.

(5) *Interruptions.* This refers to speech, laughter or actions of one member which is simultaneous with the speech of another. A family without interruptions would be reflecting some form of abnormality though not necessarily pathology. At the upper extreme, interruptions are disruptive of communication either by their intensity, timing, frequency or duration.

(6) *Laughter.* This refers to the physical act and includes giggling but not smiling. It may be an expression of fear, embarrassment, manic excitement, or fatuity as well as of good humour. The score is for frequency.

(7) *Equality of participation.* This refers to the degree to which all members are actively involved in the interview assessed both quantitatively and qualitatively. It is independent of the level of the family's overall involvement, e.g. if the family is characterized by a lack of interaction but each member contributed the same small amount of comment this would result in a high score. The involvement requires some activity and passive listening does not count. Non-disruptive play by younger children which allows them to hear is counted but older children are expected to provide more verbal contributions. Sub-grouping in which each group has full involvement of its members scores 3. A score of 1 means that at least 2 members are non-contributing in comparison to the others.

(8) *Self-affirmation.* This refers to the assertion of individuality and recognition of self by family members. At the upper levels speakers make definite statements and have clear opinions; they are able to say 'I want . . .' 'I think . . .' 'I will . . .' etc. At moderate levels commitment may be avoided on some issues or by some members and at the lower levels the family is characterized by deflection, parrying and other avoidance tactics such as silence.

(9) *Requests for commitment.* This refers to the inverse of category 8, i.e. one member's attempts to get another member to do something or to commit themselves via orders, questions, demands, requests or encouragement.

(10) & (11) *Agreement & Disagreement.* This refers to the explicitness of agreement or disagreement and is a measure of communication. The scores do not reflect whether the family do successfully agree or disagree. Explicit non-verbal communication (e.g. head-nodding) is included.

(12) *Positive support.* This refers to both verbal and non-verbal interactions between members. It includes clear evidence of understanding, affection and praise and not simply non-attacking or polite non-specific remarks and behaviour.

(13) *Attack.* This refers to both verbal and non-verbal interactions including negative attitudes, antagonism, destructive criticism, and hostile behaviour.

(14) *Intrusiveness.* This refers to inappropriate impingement on links between family members by speech or action of another member. For example, one member speaking when another has been explicitly invited to, one member relaying messages from a second to a third member, one member speaking for another member, a child interfering, without obvious cause, with inter-parent or parent–interviewer exchanges.

(15) *Mind-reading.* This refers to a phenomenon in which one member insists he knows what another is thinking or feeling in the face of evidence to the contrary including reasonable assertion from the other person. It does not include empathic comments (e.g. Mother says to a crying child 'I see you are unhappy') or faulty attempts to make sense of the behaviour of another member (e.g. Mother says of a tantrum due to some specific frustration 'He is getting bored'). A score of 5 is associated with serious distortion, denial and lying and a form of psychopathic dishonesty, e.g. a wife described the family holiday as disastrous and her husband insisted that she had really enjoyed it; or part of psychotic confusion and projective identification, e.g. a thin mother insisted that her obese daughter was hungry despite the daughter's protests that she had just had a big lunch.

(16) *Range of affects.* Affects are grouped as follows: I—fear, anxiety, helplessness, confusion. II—anger, hate, irritability, rage, jealousy, envy. III—guilt, shame, embarrassment. IV—sadness, misery, depression, despair. V—affection, love, warmth, concern. VI—pleasure, gladness, joy, pride, happiness. An affect group is judged as being represented if on at least one occasion one of the component affects is clearly displayed *and* acknowledged by at least one member of the family. To score 5, five groups must be represented; 4, four groups including V or VI; 3, four groups or three groups including V or VI; 2, three groups of I to IV; 1, two groups or less.

(17) *Intensity of affects.* This refers to the strength of feeling or intensity of expression irrespective of the range. It can vary from intense emotion which breaks through conventional social controls, such as adult crying, to intense or enthusiastic

expression of feeling, and down through low-keyed subdued expression of feeling to a withdrawn, flat or bland interview.

(18) & (19) *Tension & Comfort.* This refers to the emotional atmosphere and the amounts, respectively, of emergency emotions and welfare emotions. The rater uses verbal and non-verbal cues, and acknowledgment by the family is not relevant (cf. 16).

(20) *Humour.* This refers to the gentle irony that makes the tribulations of family life bearable. It does not include cleverness, sarcasm, or avoidance of issues by joking generalizations. It is estimated from words and tone of voice.

(21) *Effectual parental coalition.* This refers to the parental capacity to nurture and socialize children within the family and hence can be rated for a single-parent family. It does not include marital tension or discord except indirectly by interference with parenting abilities. If one parent has a preferred major coalition with a child then although parenting may be moderately successful the parental coalition is severely disrupted and the rating is 1 or 2.

(22) *Maintenance of generational boundaries.* This refers to the maintenance of the appropriate parent–child distinctions in responsibilities and roles. Both well-defined and excessively rigid boundaries score 5. Lower scores are associated with role-reversal, infantilization of parents and parentification of children, or loss of boundaries leading to a lack of clarity of role requirements, and confusion.

(23) *Alliances.* This refers to the meaningful working links between family members. All alliances will not be of equal strength or importance at a particular time but a dormant alliance should be susceptible of activation. A major split or scapegoating of one member results in a score of 1. With very young children, the amount of physical contact including restraint is taken into account.

(24) *Resonance.* This refers to the degree of reactiveness and individuation within the family. At one extreme the family is 'enmeshed': there is excessive involvement and reactiveness and the family does not behave as if it is constituted of separate individuals. At the other extreme, 'disengagement', the members do not behave as if they have feelings of loyalty or belonging and show a minimal response to the distress of others. As the raters found that both tendencies were commonly present in the same interview, they were asked to rate the preponderant tendency.

(25) *Flexibility.* This refers to the capacity of the family to reshuffle its coalitions, roles, and routines in response to changing circumstances. Persistence with habitual patterns in the face of advice, requests or confrontation as to the necessity for change, i.e. rigidity, 'high homeostasis', lowers the score. At the interview a flexible family can allow different members to be the centre of attention, can allow authority to be vested in the interviewer and can respond adaptively to difficult questions or incidents during the interview.

(26) *Conflict acknowledgment.* This refers to the family's capacity to recognize, describe and agree on the existence of specific interpersonal intrafamilial conflicts.

The intensity or distress of conflict is not relevant and a score of 5 is obtained both by healthy families aware of disagreements which are inevitable in close relationships and by families riddled with pathological conflict and aware of their overt fighting. A score of 1 is obtained if the family insists it has no conflicts in the face of specific probing for marital, parent–child and inter-sibling conflicts. Intra-psychic conflict is not relevant. (Subsequently this area was refined and an additional category of *Conflict resolution* developed.)

(27) *Feeling of safety.* This is a bi-polar scale. At one extreme is 'overprotection' rated when members are over-interfering, refusing to treat each other as adequate and usually submerged in a claustrophobic atmosphere. At the other extreme is 'neglect' with members too self-absorbed to be consistently caring about the feelings and needs of others. This is associated with an atmosphere of insecurity or danger. A family is scored at 3 when it provides adequate protection and the parents allow for age-appropriate behaviour and provide healthy inattention and non-interference.

(28) *Identity struggles.* This refers to the degree to which the family is preoccupied and fighting over what kind of person each member is. It is related to the development of autonomy and separateness as well as to identity formation within families. A low score indicates that the family is not debating self-concepts. A high score is given when there are frequent self-assertions and attributions to others with counter-assertions and counter-attributions. Members deny possession of alleged traits and may complain of not being properly recognized or understood.

(29) *Experience of the environment.* This refers to the nature of the family's experience of the environment and is determined from spontaneous comments of the family and the relationship of the family to the interviewer and his Institution. If the environment is viewed as basically helpful, intelligible and worth being connected with the score is 5. An uncertain or ambivalent relation to the environment or one in which different family members have different attitudes to a marked degree scores 3, and if the environment is experienced as hostile, confusing and best kept at bay the score is 1.

(30) *Grasp of meaning.* This is a family equivalent of psychological-mindedness and refers to the family's notion of itself as a family and whether it can understand the behaviour of its members including any symptoms as manifestations of family problems.