

across all families. Global measures of mental health in adults (Endicott, Spitzer, Fleiss & Cohen, 1976; Luborsky, 1962) and children (Shaffer et al., 1983) based on clinical judgement have proved feasible and useful. In principle, therefore, there seemed no reason why a global measure of effectiveness of family functioning or "family health" could not be devised. We have developed an instrument for making this quantitative assessment, called the Family Health Scales (FHS). (McEwan [1974] suggested the term "familial health" would be preferable to avoid confusion with measures of the physical and mental health of individuals in a family, but his term has not caught on.)

Parsons (1959) suggested that health might be defined as "the state of optimum capacity for the effective performance of valued tasks." Note the value-laden quality of family health assessment. Health does not refer to that which is normative or average, but rather, to the absence of overt pathology and to some ideal or optimal functioning over time (Offer & Sabshin, 1966). The valued tasks with which family therapists are concerned are derived from clinical psychotherapeutic practice. However, clinicians practice, and, thereby, articulate their values within the particular theoretical framework that they hold. They are reluctant to step outside the set of ideas they adhere to. These differences among clinicians pose difficulties in devising a scale. We will describe how we have handled such problems as part of our attempt to quantify the clinical valuation of family health, and will offer evidence for the reliability, validity, and applicability of our instrument.

REVIEW OF THE LITERATURE

The family therapy literature is characterized by a variety of competing theoretical schools. Despite many calls for diagnostic systemization (e.g., Bloch, 1977; DeWitt, 1978; Gurman & Kniskern, 1978) and a number of attempts to generate an encompassing framework (e.g., Fisher, 1977; Tseng & McDermott, 1978), there have been few empirical attempts to devise useful qualitative and quantitative methods of family description. The work of Reiss (1971) and Olson, Sprenkle & Russell (1979) must be noted, although their concern has been diagnostic typology rather than description of the well functioning/poorly functioning continuum. Although a large number of measurement techniques have been applied to the family life (Cromwell, Olson & Fournier, 1976; Straus, 1969), by and large they are distant from the holistic and multidimensional nature of routine clinical appraisal of families. Some relevant instruments (e.g., Behrens, Meyers, Goldfarb, Goldfarb & Fieldsteel, 1969; Epstein, Baldwin & Bishop, 1983) have been value-based and multidimensional, but have avoided global assessment.

Lewis, Beavers, Gossett & Phillips (1976) found that "raters agreed with excellent reliability about the relative health of a broad range of families" (p.199). However, their Family Health Pathology Rating Scale was used for ranking families rather than rating them. Reliability of ranking one group of families was between 0.65 and 0.90, but this contained two extreme subgroups (nonlabelled healthy volunteers and adolescent inpatient families) which might have artificially elevated the coefficient. Reliability on the second group of homogenous families was far from satisfactory: 0.28 to 0.54. (It is worth noting here, that we tried a similar exercise to the above as part of our development work: Experienced clinicians were asked to rate "overall family disorder" on 100-point scale, [0 = most pathological; 100 = most healthy]. However, we encountered intense resistance and dissatisfaction with it from our clinicians, and reliability was unsatisfactory.)

Lewis et al. (1976) also developed 13 Family Evaluation Scales (FES), each with 9 points and with anchoring definitions at 3, 4 or 5 of the points. These Scales were developed from a clinical theory of Beavers (1976), though they did not relate to it closely. The FES included: Overt Power, Parental Coalition, Closeness, Congruence with

Reality, Efficiency, Communication of Self-Concept, Responsibility, Invasiveness, Permeability, Expressiveness, Mood and Tone, Conflict, Empathy. Experienced clinicians could agree on 10 of the 13 scales (correlations: 0.57 to 0.82). When the sum of the scales were taken, raters correlated highly (0.82). Evidence for the validity of these ratings came from significant correlations with the global scale, with family members' own views of themselves and from correlation with the degree of disturbance of the adolescent index-patient.

Beavers' subsequent work (summarized in Beavers and Voeller, 1983) has moved away from the use of subscales to descriptions of families on two dimensions. One of these (somewhat mysteriously called "the negentropic dimension") is, in effect, a scale of family health with categories of severely disturbed, borderline, midrange, adequate and optimal. However, no evidence as to the reliability with which clinicians can assign families to these categories is offered.

DEVELOPMENT OF THE FAMILY HEALTH SCALES

Preliminary Work

Over the past decade, our research team has been attempting to capture the intuitive assessment made by clinicians so as to formalize it explicitly. Our initial efforts (Kinston, Loader & Stratford, 1979) were based on the collection and evaluation of commonly used categories of family interaction in the literature. An instrument, the Current Family State Assessment, was produced, containing about 30 items of interaction, each on a 5-point rating scale with anchoring definitions. Although reliability coefficients were mostly satisfactory, the instrument did not prove easy to use or suitable for our clinical research because interaction was not explicitly valued, for example, in terms of its appropriateness or disruptiveness (cf. Loader, Kinston & Stratford, 1978).

In retrospect, the main value of the CFSA may have been to convince us that a global rating was conceptually acceptable and clinically essential. Our pilot studies of the raw global scale (the 0-100 "overall family disorder" scale mentioned above) had shown us that clinicians need help in reaching a rating, and revealed a general weakness among therapists in precise description of family interaction. We were, therefore, spurred on in our efforts to develop accurate and systematic clinical descriptions of clinically elicited family interaction (Bingley, Loader & Kinston, 1984; Loader, Burck, Stratford, Kinston & Bentovim, 1981).

Aims of the Scale

Our overriding purpose was to produce an instrument which could provide a measure of family health which was clinically meaningful and would enable comparisons across families. It was not enough that the FHS should be clinically based; in addition, the process of rating should, itself, conform to clinical logic so far as possible. For example, the clinician should be allowed to use his or her ability to recognize defensive behavior, to give extra weight to pathological interactions he or she judges as particularly significant, and to set observations in the context of family size or children's ages.

We wanted the instrument to be acceptable to a wide range of clinicians working with families, irrespective of their theoretical orientation. We also wanted it to be applicable to most families without restriction, such as requiring the presence of an identified patient, or limitations on family size or other family characteristics. It was essential that the instrument be demonstrated to have adequate psychometric properties of validity and reliability. In addition, the instrument should not be too time consuming to complete.

Any method of family interaction rating depends on the means whereby interaction is elicited and brought to the rater's attention. We wanted our instrument to be usable

and reliable with any source of information, though, necessarily, the validity of any rating would be bound up with the validity of the source of information.

Structure

The Family Health Scales are constructed to yield a single figure to overall family health. The FHS consists of a number of dimensions, the "main scales," which were arrived at from reviewing the literature and from our clinical analysis of family interaction, structure and competence. The dimensions were chosen with the aim of centering ourselves in the theoretical middle ground of family therapy. We have chosen mainstream concepts of "family systems theory," which we believe are adhered to in one form or another by the majority of present-day family therapy theorists, whatever the exact details of their orientation. As we gained experience, we repeatedly revised the instrument on intuitive and empirical grounds, to maximize its usefulness and acceptability. There are 6 mainscales in the current edition of the FHS and they are labelled: Affective Status, Communication, Boundaries, Alliances, Adaptability and Stability, Family Competence. These dimensions were chosen as clinically important and distinctively recognisable. However, we expected considerable overlap amongst them and, therefore, a high degree of correlation between them and with the final FHS score.

Each mainscale is still too abstract and general to allow for a natural clinical rating. Each is, therefore, described in terms of a number of subscales, varying from 3 to 6 subscales per mainscale, with 26 in all (Table 1). Each subscale is a 7-point ordinal scale with anchoring descriptions at the first, third, fifth and seventh point. These anchors are so organized as to reflect optimal functioning at 7, adequate function at 5, significant dysfunction at 3, and breakdown of functioning at 1.

Principles of Rating

As a result of our own work and that of others, we believed that a useful score could be obtained by combining ratings made on a number of dimensions. However, we did not regard the crude "sum of all scales" used by Lewis et al. (1976) or Behrens et al. (1969) as satisfactory. Such an approach compels the rater to complete all scales and every subscale rating would contribute to the final score regardless of its clinical value. We found that in any given family, a rater might not be confident about rating a particular scale because of lack of information, or because of its theoretical implications,

Table 1
FHS Mainscales and Subscales

<i>Affective Status</i>	<i>Alliances</i>
Family atmosphere	Pattern of relationships
Nature of relationships	Marital relationship
Emotional involvement	Parental relationship
Affective expression	Parent-child relationship
Individual mood	Child-parent relationship
	Sibling relationships
<i>Communication</i>	<i>Adaptability and Stability</i>
Continuity	Family stability
Involvement	Relationship to environment
Expression of messages	Adaptability of family organization
Reception of messages	
<i>Boundaries</i>	<i>Family Competence</i>
Family's relationship to the environment	Conflict resolution
Family cohesion	Decision-making
Intergenerational boundary	Problem-solving
Individual autonomy	Parental management of children

or because he felt it did not properly fit the family under consideration. We, therefore, developed the following procedure:

The rater must consider each mainscale in the light of his total knowledge of the family. Each subscale is examined and as many as possible are rated, i.e., given a score between 1 and 7. During the rating, the rater is advised to write descriptive comments on the form to facilitate subsequent discussion and reduce the blurring of content differences between the scales. If the rater is undecided between two adjacent points, he should score towards health. The scored subscales are then considered together by the rater in the light of the definition of the dimension, and each is weighted as to its relevance and appropriateness for that particular family. A score between 1 and 7 is then decided on for that dimension (mainscale). This is the score that contributes to the FHS score, not the subscale scores; it is a judgement, and not the average of the subscales. In other words, the subscale ratings have acted as sensitizers and have served to help focus and clarify the rater's mind on the dimension in hand, and the quality of functioning of the family in that dimension. If the rater is particularly uncomfortable with any mainscale dimension, he may choose not to enter any score at all. This procedure is repeated for each mainscale. The Family Health Score is the average of the sum of as many mainscales as could be rated.

It follows from the above, that the FHS is not intended to produce a profile of scores on a variety of dimensions. Each dimension is to be seen more as a way of assessing the family as a whole, rather than as a way of measuring the particular dimension.

The rater makes his intuitive judgements holistically, that is to say "with all things considered." A focus on the purpose and effects of interaction, rather than on concrete behaviors, should enable cultural and subcultural differences amongst families to be accepted without inappropriately low ratings. However, the rater must not use subcultural norms or unfavorable social circumstances as an excuse to avoid rating dysfunction when it is present. For example, empathic and communicative dysfunction in a family suffering multiple deprivations is not to be judged as less significant of the presence of poverty or poor education.

Content of the Scales

The structure of the FHS has been provided in Table 1. The mainscales and subscales are not particularly new or original, but represent our distillation of clinical and research experience with family description and rating. A sample mainscale, Affective Status, is reproduced in Table 2.

The subscales for each dimension have been arranged, as far as possible, so that the rater may consider the family first as a family system interacting with the wider environment, then as a network of relationships, and finally as a collection of individuals. For example, if we consider the mainscale of Affective Status, then Family Atmosphere and Nature of Relationships are family-level notions. Emotional Involvement and Affective Expression refer to the network of relationships in the family, and Individual Mood corresponds to individual functioning.

An instrument like this depends on the quality of its anchor descriptions, which need to be short, pithy and easy to use. We found that concrete examples of functioning were misleading, often conflicting with clinical judgement because of the influence of context. Clearly, to provide all the concrete possibilities for which a family could be rated dysfunctional for any subscale is an impossible task; and the essence of clinical judgement is assigning a value to any particular concrete event. Such values are not usually embodied in numbers or degrees of dysfunction but in evocative words. Our approach, therefore, was to devise abstract phrases, often adjectival, to convey a notion of the continuum in each subscale. The aim of ordinal scaling is to match numbers to these words. Thus, the rater bases his rating on as much knowledge of family behavior

Table 2
The Affective Status Scale of the Family Health Scales

	Breakdown 1	2	Dysfunctional 3	4	Adequate 5	6	Optimal 7
Family Atmosphere	<input type="checkbox"/> Dead, chaotic, sense of panic, intense discomfort, claustrophobic.		Uncomfortable, cold, tense, unsafe, over-excited.		Basic sense of safety, but with some tensions.		Comfortable, vital, warm, harmonious, sense of safety, humor available.
Nature of Relationships	<input type="checkbox"/> Perverse, attacking rejecting, devaluing, overdependent.		Unsupportive, unappreciative, inconsistent, undermining.		Relationships supportive but with some inconsistencies.		Affiliative, supportive, valuing, appreciative.
Emotional Involvement	<input type="checkbox"/> Absence of involvement or intense over-involvement—positive or negative.		Detachment, over-responsiveness; emotional intrusiveness.		Attachment with marginal over- or under-involvement.		Empathic relations: Understanding without intrusion.
Affective Expression	<input type="checkbox"/> Feelings concealed or used manipulatively; expression of affect is overwhelming or absent.		Restricted range of emotions; impoverished, confusing or inconsistent expression.		Adequate expression of feelings with some difficulties.		Clear, open, direct, spontaneous and sensitive; full range of emotions available.
Individual Mood	<input type="checkbox"/> Inappropriate affect and/or painful or negative emotions predominate.		Members are ill-at-ease, flat, depressed, over-excited.		Family members reasonably at ease with themselves and their family.		Prevailing mood of members is appropriate to the situation & based in wellbeing.
MAIN-SCALE SCORE [I]	<input type="checkbox"/>		Comment on anything that particularly influenced the score for a subsection.				

as is available and judges how well this indicates the family is functioning. The descriptions are not comprehensive and terms employed at any one anchor point do not belong to that point alone.

EVALUATING THE FAMILY HEALTH SCALES

We will now indicate the variety of families which were assessed with the FHS, the different contexts of the assessment, and the evidence that was accumulated as to its psychometric properties.

Methods

Source of families. The FHS has now been used with over 200 families. Families were recruited in the course of a number of pilot and research studies, as well as during routine clinical work. The bulk of our families may be grouped as follows:

1. Families of children with physical or psychosomatic illness attending various clinics in a children's hospital: coeliac disease ($n = 36$); atopic eczema ($n = 12$); obesity ($n = 13$).
2. Families of overweight children on a general practitioner's list ($n = 11$), and at a local school ($n = 13$).
3. Families of children, not otherwise labelled, attending a local school ($n = 15$).
4. Families of children with neurotic or conduct disorder referred to a child psychiatry department in a children's hospital either from outside, or from within, the hospital ($n = 27$).
5. Families of children referred for school attendance problems to an adolescent psychiatric unit ($n = 26$); or for emotional/behavioral disorders to family therapy service ($n = 12$); or to a child sexual abuse team ($n = 12$).

Sampling methods varied according to the purpose for which the family was being rated. Consecutive attenders were approached from coeliac, eczema and obesity clinics in the hospital, an adolescent unit, and child and family psychiatric units. The nonlabelled school group was a random sample. Other groups consisted of families available at the time. Basic data on most of these families has been, or will be, provided elsewhere (Kinston and Loader, 1984, in press; Stratford, Burck & Kinston, 1982), but it is clear that we have tapped a wide variety of families. The largest research project for which FHS has been used, to date, is 3-year study of families of obese children, carried out in association with the Department of Child Health at the Institute of Child Health, University of London. Procedures and details of this study are currently being prepared for publication (Kinston, Loader & Miller, 1985).

Source of raters. The raters developing and using the instrument worked at the Hospital for Sick Children, London. Six played a significant part: 2 psychiatrists, 3 social workers, and a research assistant. In addition, the instrument has been used by other researchers at the Hospital and two groups of raters in other institutions. A large number of clinicians have tried out the scales during its development and in the course of conference demonstrations, workshops, and teaching exercises.

Source of information. Family interaction was elicited during routine diagnostic and therapeutic interviews, or by research interviews. The research interviews were either a Standardised Clinical Family Interview (SCFI) (Kinston & Loader, 1984, in press) or the Family Task Interview (FTI) (Kinston & Loader, 1986). Ratings of the interaction elicited, were made on the basis of: (a) direct contact with the family in the clinical or research setting; (b) direct observations via one-way screen, closed circuit television (CCTV), or videotape viewing; and (c) written descriptions.

The written records used for rating were of two sorts. The first consisted of 2-3 pages of free description by experienced clinicians using consecutive routine diagnostic

interviews. The second consisted of systematic descriptions with the Family Interaction Summary Format (Loader et al., 1981) by research clinicians using the Family Task Interview.

Data handling and analysis. Data was coded and punched on to computer cards and analyses performed using SPSS (Statistical Package for the Social Sciences) on the University of London Computer. Nonparametric methods of analysis such as Spearman's rank-correlation coefficient and Wilcoxon Matched Pair Signed Ranks Test were used (Siegel, 1956).

Reliability

Reliability is crucial. Inter-rater agreement for an instrument such as the FHS is a form of consensual validation. If clinicians can agree closely as to the degree of clinical disturbance, then there is no higher authority to dispute this assessment. If clinicians cannot easily agree, then serious doubt must be thrown on the whole enterprise. Intra-rater or test-retest reliability is also essential if the instrument is to be used repeatedly over a long period of time or if comparisons are to be made between groups rated at different times. If the FHS taps into a real clinical construct then intra-rater reliability should be easily demonstrable.

Agreement between raters. As the FHS has been in a developmental phase, we have routinely used two or more raters concurrently. This has given us numerous opportunities to test for interrater reliability. Table 3 lists the various studies of interrater reliability, together with the obtained correlations and tests for systematic bias. It will be seen that the correlations are almost uniformly satisfactory, mainly in the range .75-.85. Occasionally, the correlation dips down to .65-.75; the lowest value is .59. Poor reliability was usually explainable in terms of raters who were tired, bored, distracted or anxious. On a few occasions, there was a trend for one rater to be regularly rating higher or lower than the other. (Such systematic bias was regularly sought, using the Wilcoxon test.)

Test-retest reliability. Three specific studies were carried out to check for test-retest reliability. In the first, 12 videotapes of families with obese children performing the FTI were rated by two raters once, and then again, 12 or more months later. The interrater reliability in each case was .90 and .88. The intrarater reliability for each rater was: .82 and .88, and using mean FHS score on each occasion, the overall test-retest reliability was .91. In the second study, 12 families from the school group were rated by raters on the basis of the Family Task Interview and 3 months later were given another similar (but not identical) Family Task Interview. The three-way interrater reliability (Kendall's coefficient of concordance) for each interview was .74 (first) and .86 (second), and the overall test-retest reliability (calculated using mean FHS score for each occasion) was 0.79. The test-retest reliability taken rater by rater (i.e., intrarater reliability) was .80, .52, and .60. The last two values are unsatisfactory and lower than usual; they appear to be due to poorer accuracy (as manifested by reliability) on the first interview (see Table 2). In the third study, two ratings of the same FTI, given to 12 psychiatric families, were carried out within 6 months of each other by two clinicians. Interrater reliability on each occasion was satisfactory (.80 and .75) and test-retest reliability using mean scores was .91.

Validity

It is necessary to distinguish between the validity of the instrument in general, and the validity of any particular measure. This is because the confidence placed in a particular use of the instrument is dependent both on the validity of the method used for eliciting the family interaction to be rated, and on any possible distortions between the raw interaction and what is presented to the rater. Thus, for example, a particular

Table 3
Inter-rater Reliabilities of the Family Health Scales

Sample	Source of Rating	No. of Raters	Pairs	Inter-rater Reliability ^a	Systematic Bias ^b	
Psychiatric clinic <i>N</i> = 15	Free written descriptions	4	A & E	.95	N/A	
			A & F	.80	N/A	
			E & F	.88	N/A	
			A & F	.84	N/A	
			D & E	.82	N/A	
			D & F	.79	N/A	
Adolescent unit <i>N</i> = 26	Summary format descriptions	2	A & B	.83	N.S.	
Psychiatric clinic <i>N</i> = 12 (Test-retest study)	FTI (Time 1)	2	D & E	.80	.86	
	FTI (Time 2)	2	D & E	.75	.93	
Coeliac clinic A: <i>N</i> = 15 B: <i>N</i> = 17 C: <i>N</i> = 12	SCFI	2	A & D	.73	.51	
			E & F	.37**	.09*	
			A & B	.80	.42	
	FTI	3	A & C	.77	.41	
			B & C	.75	.54	
			A & B	.87	.08*	
D: <i>N</i> = 13	FTI	3	A & C	.74	.53	
			B & C	.82	.42	
			A & B	.80	.42	
Mixed group <i>N</i> = 12	FTI	3	A & C	.77	.41	
			B & C	.75	.54	
			A & B	.80	.42	
Obesity group A: <i>N</i> = 37	FTI	3	A & B	.87	.84	
			A & C	.69	.17	
			B & C	.65	.22	
	B: <i>N</i> = 12 (Test-retest study)	FTI (Time 1)	2	A & B	.90	.80
		FTI (Time 2)	2	A & B	.88	.52
	School Families <i>N</i> = 15 (Test-retest study)	FTI-A	3	A & B	.63	.80*
A & C				.59*	.29	
B & C				.59*	.23	
FTI-B		3	A & B	.77	.92	
			A & C	.88	.53	
			B & C	.70	.76	

^aReliability measured nonparametrically with Spearman's rank correlation coefficient using rater pairs.

^bLikelihood of systematic bias is checked by Wilcoxon Matched Pairs Signed Ranks Test; *p* values are listed; N/A = test not performed; N.S. = not statistically significant.

*Probably unsatisfactory. **Seriously unsatisfactory.

measure may be invalid either because the interviewer did not permit the family to show its pathology, or because the descriptions of adequately elicited interaction were biased. With regard to validity of the instrument as such, there are a variety of inter-linked questions to be answered: Does the scale adequately represent the domain it is supposed to measure? Does the score correspond to some other, more accurate, measure of family functioning? Does the score specifically measure family health or merely some

closely related construct such as the mental health of the sickest member? Does the score alter in line with improvement or deterioration of the family?

Content validity. The presupposition underlying the construction of the FHS is that different clinicians and different theories tend to concentrate on different aspects of family interaction. The six main dimensions are each regarded as capable, alone, of providing a measure of family health. Each mainscale measure, therefore, acts as a form of cross-check for the others, and so the mean of mainscale scores should be the best estimate of family health. An implication of this model is that the various mainscales for each family should be very similar, and that the mainscale scores should show a very high positive correlation with the FHS score.

Table 4 shows typical mainscale score ranges and their correlations with the FHS score in the obesity study ($N = 67$). Almost all correlations are between 0.85 and 0.95, the lowest being .77. Similar results have been obtained in the numerous studies we have carried out, and also apply when subgroups of families are studied.

If the FHS is tapping into the overall quality of family functioning, it should be being used over much of its range. Table 5 lists typical ranges of scores for six raters on several groups of families with information on interaction obtained in different ways. It is apparent that scores at both ends of the range are not uncommon in all the studies.

Discriminative validity. For the scale to be valid it must, first of all, be able to discriminate between groups of families whose levels of functioning are intuitively known to be different. We predicted that the FHS means and standard deviations would be similar for groups obtained through some form of psychiatric referral and similar for groups generated by a nonlabelled or physically ill member, but that these two types of families would be discriminated. Table 5 bears this out, with psychiatrically labelled families having mean scores between 3 and 4, and nonlabelled families having mean scores between 4 and 5. Statistical calculations have not been provided in this case because first, no deliberate study of discriminative validity was performed, second, the raters were not always blind to the group, and third, the methods of obtaining FHS scores are not comparable. Despite the evidence for discrimination, examination of the standard deviations suggests that there is a considerable degree of overlap.

Table 4
Subscale Score Ranges and Correlations with FHS Scores

Main Scale	Rater	Range Used	Correlation with FHS
Affective Status	A	1-7	.86
	B	2-7	.94
	C	2-6	.92
Communication	A	2-6	.88
	B	2-6	.92
	C	2-7	.87
Boundaries	A	2-7	.77
	B	2-7	.87
	C	1-7	.88
Alliances	A	2-7	.87
	B	2-7	.87
	C	2-6	.88
Adaptability and Stability	A	2-7	.91
	B	2-6	.86
	C	1-7	.87
Family Competence	A	2-6	.85
	B	3-7	.83
	C	2-7	.90

Table 5
Use of the Family Health Scales in a Variety of Studies by a Variety of Raters.
Means, Standard Deviations and Ranges of FHS Scores With Seven Different
Samples

Study or Sample Source of Information	N	Rater	Range	Mean (SD)
Psychiatric referrals (written descriptions of diagnostic interviews)	15	A	1.7-5.0	3.1 (1.0)
		D	1.3-4.7	2.8 (.98)
		E	2.0-5.0	3.6 (.90)
		F	1.0-4.4	2.7 (1.0)
Psychiatric clinic (FTI)	12	D	1.4-5.0	3.5 (1.0)
		E	1.6-5.7	4.0 (1.3)
School Non-attenders (FTI using Summary Format Reports)	26	A	2.5-6.2	3.5 (1.0)
		B	1.5-5.8	3.4 (1.0)
Coeliac Clinic (1979 Sample) (SCFI)	15	A	2.9-6.0	4.8 (.98)
		D	3.7-5.7	4.7 (.74)
(FTI)	17	E	3.6-6.1	5.1 (.72)
		F	2.4-6.6	4.6 (1.2)
(Re-rating of FTI)	12	A	3.5-6.0	4.9 (.80)
		B	3.8-6.0	4.8 (.69)
Coeliac Clinic (1981 Sample) (FTI)	13	A	2.7-5.3	4.1 (0.8)
		B	2.5-5.7	4.4 (0.9)
		C	3.2-5.5	4.3 (0.8)
Random local school sample (FTI)	15	A	2.2-6.7	4.2 (1.2)
		B	2.3-6.7	4.6 (1.1)
		C	2.8-6.5	4.3 (1.1)
Obese sample (FTI) (from hospital clinic, GP practice and local school)	37	A	2.7-5.7	4.2 (0.8)
		B	2.0-6.0	4.2 (1.1)
		C	2.0-6.0	4.0 (1.1)

Sensitivity. The above results are evidence for validity, but of a very crude sort. The instrument was not devised simply to discriminate between groups but also within them. This raises the question of sensitivity. Is the FHS sensitive to the expected wide variation in family functioning within any particular group, and can it identify changes in family health over time? Some evidence for sensitivity can be provided by looking at the distribution of scores within groups, like the random sample from school or the families with a physically ill child, where it should be expected that a substantial portion of the scale would be used. Table 5 indicates that this, indeed, is the case.

Specificity

Comparison of the FHS with a more accurate measure of family health is not possible because there is no such accepted accurate measure. Hence, comparison with other indicators of family health is better regarded as a test for specificity than for validity, as such. In other words, it is important to ask whether the FHS taps into something which is distinctly different from other assessments of family life. Those which we have considered include self-report assessments of the family, individual psychological health, and other psychosocial factors relevant to family life.

Self-report assessment of family life. It is possible to obtain a measure of family health by a self-report technique. However, such a measure made by a family insider would be expected to differ from that made by an outsider, especially a family therapist outsider, and there is evidence to support this (Olivieri & Reiss, 1984; Sigafos, Reiss,

Rich & Douglas, 1985). Correlations would be expected to vary depending on the level of openness or self-awareness of the family—rarely moderately positive, usually low positive, and sometimes negative for defensive individuals. Such a test was carried out using the Family Functioning Index (Pless & Satterwhite, 1973) which is completed by the two parents separately. Details of the study are reported in Kinston et al. (1985). The three groups used were: (a) families of obese children recruited from the hospital (hospital obese), (b) families of overweight children recruited from general practice (community obese), and (c) families of children with coeliac disease attending for routine check-ups at the hospital (hospital coeliac). The results are tabulated in Table 6.

Treating the families together as one large group revealed virtually no correlation. Taking the families group by group, revealed the expected variation from low negative to low positive correlations, depending on the group and on which parent is acting as reporter on the family.

Individual psychopathology. The relationship between individual disturbance and family dysfunction is not well understood. The presence of an ill member is often said to betoken family pathology, but whether such member dysfunction correlates with family dysfunction as measured by the assessment of family interaction is unknown. We might expect high correlations at the extremes—optimally functioning families to contain well functioning members, and severely dysfunctional families to be constituted with severely disturbed members. However, in the mid-range where the majority of families lie, a lack of congruence between family health and individual health would be more likely. In other words, we would predict that, in general, an FHS assessment could not simply be replaced by assessments of individual mental health.

We have examined this possibility in several studies in which FHS assessment was based on interaction generated by the SCFI or (more usually) the FTI. At the conclusion of each standardized interview, we administered well established, self-report questionnaires to get an estimate of individual psychological health. The parents completed the Rutter A Scale (PR) for school children (Rutter, Tizard & Whitmore, 1970) and the Behaviour Check List (BCL) for pre-school children (Richman, 1977). Both parents and any children 16 years or over completed the 60-item General Health Questionnaire (GHQ) (Goldberg, 1972). All three instruments employ a cut-off score, above which the individual concerned is deemed a potential psychiatric case. The GHQ has further properties: The higher the score, the greater the probability of being a case, and the more severe the case. It should be noted that though these instruments are widely used (in the United Kingdom at least), none have the specificity and sensitivity of a clinical-research interview. The family was deemed unhealthy on the individual health criterion if one member or more scored as a "case" on the GHQ, PR or BCL.

The unhealthy/healthy cut-off chosen for FHS scores was 4.4/4.5. This value was chosen, first, on theoretical grounds: The scale construction allowed for a score of 5 to indicate adequate functioning, and a half-point is regarded as the maximum likely error in rating; and, second, on empirical grounds: We have found that the mean FHS scores of groups of nonlabelled families usually lie between 4 and 5 (cf. Table 5).

Table 6
Correlations Between FHS and FFI Scores

Comparison	Hospital Obese <i>n</i> = 13	Community Obese <i>n</i> = 9	Hospital Coeliac <i>n</i> = 11	All Families <i>n</i> = 34
Mother's FFI and FHS	-.20	.46	.16	-.02
Father's FFI and FHS	-.29	.04	.30	.09

FHS ratings were always made in ignorance of scores on the instruments measuring individual health. We then compared these two indicators of family health. Congruence between them meant that families were rated as healthy or unhealthy on both criteria. Incongruence meant being assigned healthy by one criterion and unhealthy by the other. It seemed likely that the degree of congruence might vary in the different groups of families, either due to the degree or type of pathology, or because of the desire of family members to present themselves as healthy on self-report.

Five different groups collected in research studies were examined: (a) 15 coeliac clinic families (1979 sample); (b) 12 psychiatric clinic families; (c) 13 coeliac clinic families (1981 sample); (d) 37 families with an obese child; (e) 15 randomly selected school families. The results for the five groups are displayed in Table 7. As predicted, there was marked incongruence which varied considerably—from 16% in the school group, to 32% and 38% in the coeliac groups, to 41% in the psychiatric group, to 51% in the obese group. The amount of incongruence due to ratings of unhealthy on FHS but healthy by individual case, was equal to that due to those rated healthy on the FHS but unhealthy by individual case.

It is possible that the FHS score might be closely related to individual ill health if the individual scores were treated as a continuum, or if scores were combined in some way (e.g., both parents but no children). Table 8 shows the results of a correlation analysis. The persistent weak trend for severity of individual psychopathology to be associated positively with family dysfunction is evidence for both the validity and specificity of the FHS.

Other psycho-social factors. We have also investigated the possibility that the FHS score might simply be a reflection of some other variable, such as social class. On the sample of 67 families in the obesity project, we used a One-way Analysis of Variance and found no significant relation of FHS with social class, number of children, number of marriages, one or two parent families, parental health, parental nationality, religion, amount of contact with families of origin, housing and stage of family life cycle.

Table 7
Relation Between Family Health Measured on the Family Interaction
Criterion and the Individual Health Criterion

Group	Matrix	No. of Families	% Incongruence*
Group A: Coeliac Group (1979)(SCFD)	8 2 3 2	<i>n</i> = 15	33%
Group B: Psychiatric Group (FTD)	0 3 2 7	<i>n</i> = 12	41%
Group C: Coeliac Group (1981)(FTD)	2 3 2 6	<i>n</i> = 13	38%
Group D: Obese Group (FTD)	3 9 10 15	<i>n</i> = 37	51%
Group E: School Group (FTD)	5 1 1 8	<i>n</i> = 15	16%

*Note. Incongruent = Healthy on FHS and Unhealthy Members, or, Unhealthy on FHS and Healthy Members; Congruent = Healthy on FHS and Healthy Members, or, Unhealthy on FHS and Unhealthy Members.

Table 8
Relation between FHS and Indices of Severity of Psychopathology in Individual Members of the Family

Relation	<i>N</i>	<i>r</i>	<i>p</i>
FHS and father's GHQ	57	-.21	.06
FHS and mother's GHQ	65	-.11	.19
FHS and mean parental GHQ	67	-.20	.06
FHS and index child PR	66	-.09	.25
FHS and sibling PR	41	-.07	.33
FHS and mean children's PR	67	-.22	.04

**Note.* Data from obesity study: Families included hospital and community obese, coeliac, and school nonlabelled—*N* = 67. *N* of each correlation varies because some families lack certain members. Correlations of the subgroups varied but were not appreciably different from the overall results. Health is indicated by high scores on the FHS and low scores on the other instruments. Hence, negative correlations indicate a positive association between family and individual health.

Robustness

If the FHS is to be widely used, it must be robust. If, however, FHS is sensitive to circumstances under which ratings are made, then this should be known and taken into consideration.

Duration of assessment. In our experience, it is difficult to rate the FHS with confidence after viewing a single 10-minute interaction task. The clinical style of observing is based on developing a cumulative impression of the family. Observing a repetition of pathological types of interaction is essential, otherwise the rater tends to "give the family the benefit of the doubt." We have found that a minimum of 20–30 minutes of interaction is necessary for rating. In one study, 12 families were rated based on viewing 3 of the 6 FTI tasks, with an inter-rater reliability of .80.

Set of the rater. In practice, the FHS rater is either the family interviewer or an observer. The mental set of a person in either of these two positions would naturally vary, and we were, therefore, concerned as to the effect of this mental set on ratings. (It should be noted that the observer is never in the room with the family or the interviewer. Also, the two raters often rated at different times or in different rooms. When raters were rating in the same room, there was no discussion and they sat so as to minimize nonverbal cues.)

Two studies threw some light on this issue of mental set. In the principal study, the FHS ratings of the interviewer who administered our research interview (the SCFI) were compared with those of an observer watching the interview simultaneously on CCTV. The inter-rater correlation was .75 ($p < .05$) indicating similar views of the family interaction from these two positions.

In the other study we administered the FTI to 12 clinic families, 6 provided by each of two family therapists. We compared FHS ratings made by the clinician from her role as therapist to those made by the other clinician observing the FTI on videotape. The correlation between these two scores at the two events was .59 ($p < .05$). We also compared ratings for each clinician made from her role as therapist to those made by her on the basis of FTI observation. Both clinicians showed consistent differences in ratings made

in these two situations: One of them systematically rated her families as more dysfunctional when an observer, while the other systematically rated her families as less dysfunctional. The effect of set on ratings is confounded here by the fact that the two sets are associated with different information about the family.

Degree of training. General experience and training as a family therapist is essential for use of the scales. Formal studies have not been carried out, but on several occasions, untrained personnel, knowledgeable about emotional or psychological functioning, have observed videotapes and attempted to rate. They find it difficult to do, as they are unused to the type of observation required. Experienced family therapists take to the scales reasonably easily, but a course of training in their use is necessary if they are to obtain reliability consistently.

DISCUSSION

FHS ratings have been made by two or more raters on over 200 families of varying types in varying situations over the past 7 years. It is necessary, first, to review the methods and psychometric findings, then to discuss the practicality of the instrument, and, finally, to assess its usefulness.

Methodological Issues

The FHS provides a measure of overall family health based on clinical judgement. Because a wide variety of interaction patterns are possible at any scale point, two new techniques of rating have been introduced to enhance the quality of clinical judgement. First, subscales are rated to sensitize the rater to each mainscale dimension. Second, subscales are weighted as to their significance before the rater decides on a mainscale rating. These techniques work but their mathematical and statistical implications have not been investigated. We believe that the new techniques have been crucial in producing a degree of reliability and validity that is greater than that achieved in comparable instruments. With this in mind, the psychometric findings can be looked at in more detail.

In over a dozen studies, carried out in both clinical and research settings, and using different methods of elicitation, inter-rater agreement has been well within the acceptable range. The context in which reliability has been most extensively tested is that of the Family Task Interview. Often, the reliability has been high, and only rarely has it been unacceptable. Problems with reliability have seemed to be primarily a function of poor organization of the rating process rather than due to the instrument itself. Many of the subgroups of families have been small (often around 12); however, this would have been expected to militate against achieving reliability. Checks of test-retest reliability and intra-rater reliability have also been satisfactory.

Validity is somewhat less well-established. As conventionally understood, little has been done to establish content validity, and what has been done might be criticized. For example, the high correlations between mainscale scores and the FHS scores could be explained as a halo effect. However, in the novel approach adopted where the rater has permission to omit mainscales from his rating, such halo effects are both expected and desirable. Similarly, we have only a very limited concern to establish construct validity for the mainscales, and we would strongly advise researchers not to attempt to use the mainscales as valid measures of Communication, Affective Status and so on. More research work does need to be done on discriminative validity of the FHS score, and on the construct of "overall family health," itself. Evidence has been presented that the Family Health Scales produces a score which indicates the family's overall quality of functioning as seen from a clinical perspective, and is not a reflection of some other factor like individual psychological illness. Further evidence will be available from a

study currently underway to check alterations in FHS before and after therapy, and a study of FHS change in relation to life-crisis and life-cycle change would be useful.

Our emphasis on designing the FHS as a "clinical measure" means that the highest form of validation is consensual, that is to say, agreement on ratings by trained family therapists. However, because the Family Health Scales reflects the systematic use of the "clinical eye," the full significance of measures derived from its application to families not referred for emotional or behavior disturbance is yet to be determined. The scale has been constructed so that some degree of family dysfunction is seen as compatible with effective and culturally acceptable functioning. At the lowest scores, the family is highly dysfunctional and probably very damaging to its members; at mid-levels the family has significant dysfunction whilst higher scores reflect adequate and optimal functioning. It is noteworthy that this use of 7 points with 4 anchor descriptions is similar to that produced by Beavers (1976). Informal contacts with many family therapists have also indicated that our descriptions and form of scaling do fit their conceptions of gradations in overall family health; but more systematic studies would be desirable.

Validity of the instrument-in-use is crucially dependent on the quality of the information used for rating. Ratings may be made on information about the family from a variety of sources, but ideally, should be derived from direct observation of the family. Our research experience (Kinston & Loader, 1984, 1986, in press) has led us to use the Family Task Interview (FTI) for eliciting family interaction. Videotaping of the FTI is useful, so that doubts or issues in relation to reliability and validity of a rating can be handled by reviewing the videotape.

Practical Problems

The scales have been devised to lie in the mainstream of family systems theory and to be compatible with the variety of theoretical orientations that clinicians hold. Informal trials of the instrument have revealed this to be the case in practice. The instrument is simply designed and is reported as convenient and comfortable to score. Rates complete scoring quickly—usually in 15–20 minutes and rarely over 30 minutes. A small number of clinicians have refused to rate, or have found scales confusing. However, the vast majority find rating a natural and acceptable activity and usually venture ratings on all the main scales.

Clinicians had two sorts of difficulty: first, the lack of a common terminology for family description; and second, problems in focusing on observable family interaction with enough care and without personal bias. Our research group has put considerable effort into clarifying terminology and systematizing description (Bingley et al., 1984; Kinston et al., 1979; Kinston, Bentovim & Loader, 1982; Loader et al., 1981). Clinicians are used to using terms loosely and as needed for therapeutic purposes, but using the FHS requires a certain discipline and a determination to refrain from homing in exclusively on dysfunction. The principal training called for involves developing the capacity for careful observation, recognition, and interpretation of family behavior in a setting where clinical intervention may not be required or possible. Given such a capacity, practice with the instrument is reasonably straightforward. Discussion between raters using standard, frequently rated FTI videotapes enables raters to achieve a common understanding of terms and benchmarks for scoring family dysfunction. More sophisticated training methods for raters are currently under development.

Applications

The Family Health Scales has been useful for both research and teaching. The instrument has proved valuable in promoting careful observation of family behavior, and provides a framework for the comprehensive and systematic assessment of family

interaction. In this, it serves as a companion to the Summary Format of Family Functioning (Bingley et al., 1984; Loader et al., 1981), which is a method for structuring the description of family interaction. Comparing families with the aid of the Family Health Scales also assists the trainee in articulating differences between families. We have presented our methods on many occasions, nationally and internationally, to some hundreds of family therapists, usually giving therapists an opportunity to try rating. These experiences have not only confirmed, in a general way, the validity and acceptability of the scales (as mentioned above), but have also revealed that family therapists find that the exercise of rating sharpens their observation.

However, the main use of the FHS is as a research tool. It may serve three different purposes. First, it allows hypotheses about family dysfunction to be tested. For example, in our comparison of individual and family pathology, we were examining the theoretical hypothesis of the family systems model: that events at one level of organization do not predict events at a higher level. This hypothesis was only partially confirmed. In our study of childhood obesity (to be published), we have tested the clinical hypothesis that such families were more disturbed than other families seen by physicians. This hypothesis was refuted. Second, the FHS may be used to check changes in the family over time. For example, the FHS should contribute to outcome assessment. Such a study of family therapy is now underway. This uses the FHS as well as clinical judgement and self-report judgements to assess change; and it is expected that the study will also contribute to further understanding of the properties of the FHS. Third, the FHS may be used to classify or stratify families by degree of disturbance, as part of other research studies. For example, in much child psychiatry research, if two groups are to be accepted as comparable, similar degrees of family dysfunction should be demonstrated in each group. Such comparability often appears to be more significant than the usually accepted checks such as social class, parental age and so on.

The Family Health Scales might be deemed to suffer from a "floor" problem. Its development has taken place within a context of relatively healthy families, and ratings require some form of recognizable family life and member interaction. There appears, however, to be an extensive continuum of severely deteriorated and disintegrated families that exist at score 1 and are picked up by welfare and policing agencies. An example from the lower end might be a family represented by a single adolescent without known parents and multiple care-takers and agency involvements. Family therapists do work with these individuals, although the term "family" is usually replaced by the term "client-professional system."

CONCLUSION

An instrument only achieves substantial validation over a long period of time after extensive use by many researchers in a variety of contexts with different sorts of families. In making the Family Health Scales available at this point in its development, we are expressing confidence that it is a tool with a reasonable degree of reliability and validity already, and deserving of further development and use.

REFERENCES

- Beavers, W. R. (1976). A theoretical basis for family evaluation. In J. M. Lewis, *No single thread: Psychological health in family systems*. New York: Brunner/Mazel.
- Beavers, W. R. & Voeller, M. N. (1983). Comparing and contrasting the Olson Circumplex model with the Beavers systems model. *Family Process*, 22, 85-98.
- Behrens, M. L., Meyers, D. I., Goldfarb, W., Goldbarb, N., & Fieldsteel, N. D. (1969). The Henry Ittleson Center Family Interaction Scales. *Genetic Psychology Monograph*, 80, 203-295.

- Bingley, L., Loader, P. & Kinston, W. (1984). Research Report: Further developments of a format for family description. *Australian Journal of Family Therapy*, 5, 215-218.
- Bloch, D. (1977). Notes and comments. *Family Process*, 16, 511-514.
- Cromwell, R. E., Olson, D. H. & Fournier, D. G. (1976). Tools and techniques for diagnosis and evaluation in marital and family therapy. *Family Process*, 15, 1-49.
- DeWitt, K. N. (1978). The effectiveness of family therapy: A review of outcome research. *Archives of General Psychiatry*, 35, 549-561.
- Endicott, J., Spitzer, R. L., Fleiss, J. & Cohen, J. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, 33, 766-771.
- Epstein, N. B., Baldwin, L. M. & Bishop, D. S. (1983). The McMaster Family Assessment Device. *Journal of Marital and Family Therapy*, 9, 171-180.
- Fisher, L. (1977). On the classification of families: A progress report. *Archives of General Psychiatry*, 34, 424-433.
- Furniss, T., Bentovim, A. & Kinston, W. (1983). Clinical process recording in focal family therapy. *Journal of Marital and Family Therapy*, 9, 147-170.
- Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire. Maudsley Monograph No. 21*. London: Oxford University Press.
- Gurman, A. S. & Kniskern, D. P. (1978). Research on marital and family therapy: Progress, perspective and prospect. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behaviour change: An empirical analysis* (2nd ed.). New York: Wiley.
- Kinston, W. & 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.
- Kinston, W., Bentovim, A. & Loader, P. (1982). Constructing a focal formulation and hypothesis in family therapy. *Australian Journal of Family Therapy*, 4, 37-50.
- Kinston, W. & Loader, P. (1984). Eliciting whole family interaction with a standardised clinical interview. *Journal of Family Therapy*, 6, 347-363.
- Kinston, W. & Loader, P. (1986). The Family Task Interview. Manuscript submitted for publication.
- Kinston, W. & Loader, P. (in press). Preliminary psychometric evaluation of a standardized clinical family interview. *Journal of Family Therapy*.
- Kinston, W., Loader, P. & Miller, L. (1985). *Obesity as an expression of family disturbance*. Unpublished manuscript.
- Kinston, W., Loader P. & Stratford, J. (1979). Clinical assessment of family interaction: A reliability study. *Journal of Family Therapy*, 1, 291-312.
- Lewis, J. M., Beavers, W. R. Gossett, J. T. & Phillips, V. A. (1976). *No single thread: Psychological health in family systems*. London: Macmillan.
- Loader, P., Burck, C., Stratford, J., Kinston, W. & Bentovim, W. (1981). A method for organising the clinical description of family interaction: The "Family Interaction Summary Format." *Australian Journal of Family Therapy*, 2, 131-141.
- Loader, P., Kinston, W. & Stratford, J. (1978). Is there a "psychosomatogenic" family? *Journal of Family Therapy*, 2, 311-326.
- Luborsky, L. (1962). Clinician's judgements of mental health: A proposed scale. *Archives of General Psychiatry*, 7, 407-417.
- McEwan, P. J. M. (1974). The social approach to family health studies. *Social Science and Medicine*, 8, 487-493.
- Offer, D. & Sabshin, M. (1966). *Normality*. New York: Basic Books.
- Olivieri, M. E. & Reiss, D. (1984). Family concepts and their measurement: Things are seldom what they seem. *Family Process*, 23, 33-48.
- Olson, D. H., Sprenkle, D. H. & Russell, C. S. (1979). Circumplex model of marital and family systems: I. Cohesion and adaptability dimensions, family types and clinical applications. *Family Process*, 18, 3-28.
- Parsons, T. (1959). The social structure of the family. In R. N. Anshen (Ed.). *The family: Its function and destiny*. New York: Harper.
- Pless, I. B. & Satterwhite, B. (1973). A measure of family functioning and its application. *Social Science and Medicine*, 7, 613-621.

- Reiss, D. (1971). Varieties of consensual experience, I. A theory for relating family interaction to individual thinking. *Family Process*, 10, 1-27.
- Richman, N. (1977). Is a behaviour checklist for pre-school children useful? In P. J. Graham (Ed.), *Epidemiological approaches to child psychiatry*. London: Academic Press.
- Rutter, M., Tizard, J. & Whitmore, K. (1970). *Education, health and behaviour*. London: Longmans.
- Shaffer, D.; Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H. & Aluwahlia, S. (1983). A Children's Global Assessment Scale (CGAS). *Archives of General Psychiatry*, 40, 1228-1231.
- Siegel, S. (1956). *Non-parametric statistics: For the behavioral sciences*. New York: McGraw-Hill.
- Sigafoos, A., Reiss, D., Rich, J. & Douglas, E. (1985). Pragmatics in the measurement of family functioning: An interpretive framework for methodology. *Family Process*, 24, 189-203.
- Stratford, J., Burck, C. & Kinston, W. (1982). Influence of context on the assessment of family interaction: A clinical study. *Journal of Family Therapy*, 4, 359-371.
- Straus, M. A. (1969). *Family measurement techniques: Abstracts of published instruments, 1935-1965*. Minneapolis: University of Minnesota Press.
- Tseng, W-S & McDermott, J. F. (1978). Triaxial family classification: A proposal. *Journal of the American Academy of Child Psychiatry*, 18, 22-43.

NOTE

¹Copies of the Family Health Scales (4th Edition), together with brief guidelines for training, are available at cost. For practice in rating, a sample videotape containing 3 families given the Family Task Interview is also available. For details contact Dr. W. Kinston at Brunel University, Uxbridge, Middlesex UB8 3PH, England.