

The Marital Communication Rating Schedule: An Instrument for Clinical Assessment

Joyce Borkin,^{1,4} Edwin J. Thomas,² and Claude L. Walter³

Accepted June 2, 1980

The Marital Communication Rating Schedule (MCRaS) is presented as an observationally based clinical rating system for assessing verbal behavior in marital communication. Data from 35 response display discussions lasting from 20 to 30 min each, which took place between 11 married couples, were used to examine aspects of the reliability and validity of the instrument. Three raters made independent ratings of 37 MCRaS categories for each husband and wife for each discussion period. Reliability among the raters was shown to be high when calculated within one scale point. Concurrent validity was assessed by comparing MCRaS ratings for four categories with observationally based validation criteria independently coded and measured. Results indicated that for three categories – negative statements, overgeneralizations, and amount of talk – ratings produced results that were similar to those yielded by laborious coding of audiotapes. For one category, opinions requested, a relationship between the ratings and coded data was not found. The validation results were discussed in terms of possible differences in the basis of ratings for the categories subjected to validation. Although further research is needed, it was concluded that MCRaS has many of the desirable qualities needed in a clinically useful, observationally based rating system.

KEY WORDS: marital communication; verbal behavior; communication rating.

This investigation was conducted in connection with the Sociobehavioral Research Project at The University of Michigan when Joyce Borkin and Claude L. Walter were affiliated with the project.

¹ University of Illinois, Urbana-Champaign.

² The University of Michigan.

³ Washington University.

⁴ Address all correspondence to Joyce Borkin, School of Social Work, University of Illinois at Urbana-Champaign, 1207 West Oregon Street, Urbana, Illinois 61801.

In recent years, there has been a marked increase in research in the area of marital communication, and there is now a variety of procedures to assist researchers and practitioners in the assessment of disordered marital communication (e.g., see Gottman *et al.*, 1976; Birchler, 1979; Jacobson & Martin, 1976; Thomas, 1977; Weiss & Birchler, 1978; Weiss & Margolin, 1977). However, despite the increased availability of assessment procedures, there is still the need for an observationally based instrument to assess marital communication that is practical for use in clinical work.

Most existing assessment instruments were developed for nonclinical purposes (e.g., Strodbeck, 1951; Olson & Straus, 1972). Many of the instruments have been employed in research and have necessitated detailed and time consuming postinteraction coding by highly trained coders (e.g., Carter & Thomas, 1973; Eisler *et al.*, 1973; Hops *et al.*, 1972). The Marital Interaction Coding System (MICS), for example, calls for the coding of videotaped data sequentially in 30-s intervals, using 29 code categories. This coding system has been employed in a number of important studies of marital conflict and accord. However, as Birchler (1979) has indicated, the assets of sophistication and complexity of the MICS methodology become liabilities when one considers the fact that few investigators and few clinicians command the resources to implement the procedures. Indeed, the time required to replay tapes for postsession coding could itself be a major deterrent to clinical use. Additionally, Jacobson *et al.* (1981) have pointed out that despite the relatively large amount of research done with MICS, questions concerning both its reliability and validity remain unanswered.

A clinically useful assessment instrument should ideally capture essential aspects of the highly diverse and complex verbal behavior in communication, yield clinically relevant information, and its use should be compatible with the practicalities of the typical clinical setting. More specifically, such an instrument should provide for rapid identification and selection of specific verbal behaviors for modification from among a broad band of relevant communication categories, not require extensive postsession effort, not require specialized personnel or complicated or expensive equipment, be relatively easy to learn to use, and still be sufficiently reliable and valid for clinical purposes. This is a tall order, especially inasmuch as some of these desirable characteristics may not be entirely compatible (e.g., data gathered quickly, when observationally based, may be obtained at the expense of reliability and validity). At this stage of measurement of verbal behavior in family interaction, there may be no ideal solution. Rather, each instrument may be found to have its own pattern of strengths and limitations. An important task for researchers in this domain of assessment is therefore to examine different approaches to the measurement of marital communication, giving particular attention to the possible strengths and limitations of the instruments for clinical use.

The purpose of this paper is to present findings on the reliability and validity of an observationally based rating system for assessing a broad range of

potential problems of verbal behavior in marital communication. The instrument is intended primarily to provide for rapid clinical assessment, and it may also be useful in clinical research. It is called the Marital Communication Rating Schedule (MCRaS), a revision and extension of the Verbal Problem Checklist (VPC) reported initially by Thomas *et al.* (1974) and described further by Thomas (1977). Like its predecessor, MCRaS is designed as a clinical tool to enable a therapist to assess rapidly and efficiently a large number of potentially problematic verbal response classes. It consists of 37 potentially problematic categories of verbal responses, which are rated by the clinician immediately after listening to a discussion of the marital partners obtained in the form of a "response display."

THE MARITAL COMMUNICATION RATING SCHEDULE (MCRaS)

While retaining the same purpose and many of the features of its predecessor schedule, the MCRaS was designed to increase its clarity, conciseness, and efficiency as an assessment instrument.

Rating Procedure

Each interactant is rated on each of 37 categories, given below, immediately following a response display discussion. The first 20 categories are rated on a seven-point bipolar scale. These are all classes of behavior that may be appropriate or inappropriate. The scale is identical for the first 12 items. Items 13–20 have a similar seven-point scale where 0, the midpoint, indicates appropriateness, but the poles are defined for each specific category on the scale; for example, slow to fast, quiet to loud. The label for each bipolar response category thus cues the rater to a response dimension for which judgments may be made in the direction of deficiency or excess.

The remaining 17 items are on a four-point scale (one-half of the bipolar scale) from 0 (does not occur) to +3 (occurs frequently). These categories all pertain to behaviors that are considered dysfunctional and therefore the negative side of the scale is omitted. For example, one cannot be deficient in "misrepresentation of facts" or have too few "overgeneralizations."

The Categories

Essentially all of the content of the 49 original VPC response categories was retained in MCRaS without loss of behavioral specificity. This was accomplished by combining related categories into single bipolar categories. For example, overtalk and undertalk each appeared on the VPC and were combined

on the MCRaS into one category, amount of talk, that can be rated from extreme undertalk to extreme overtalk (see category 14, below). The creation of bipolar categories in this manner decreased and consolidated the judgements required of raters without loss of information. By creating bipolar categories in this fashion, the number of response categories was reduced to 37 in contrast to the 49 of the earlier VPC.

The 37 categories on the MCRaS can be grouped into four different types of responses. The first and largest group are categories that pertain to the actual *content of the conversation*, such as positive statements (1) and negative statements (2). Others in this group are 3, 4, 7, 8, 9, 11, 21, 22, 23, 24, 26, 27, 31, and 32. The second group of categories are those primarily or exclusively concerned with the *vocal characteristics* of speech, including speed and volume (categories 14–20 and 28). A third type concerns those verbal responses that serve to *control or focus* the conversation, such as opinions requested (5) and information requested (6). Others of this type are 12, 13, and 29. Finally, categories that deal primarily with *referent representation* may be delineated, such as specific talk (10), overgeneralizations (30), and others (33, 34, 35, 36, and 37).

The rating categories and definitions are given below.

1. *Positive Statements.* A positive statement is defined as an evaluative statement indicating that something is good about the other interactant as a person or about the content of the discussion at hand. A deficiency of positive statements exists when an interactant fails to make positive statements when they appear to be called for in the discussion. An excess of positive statements exists when an interactant makes positive statements inappropriate to the content or in excessive number.

2. *Negative Statements.* A negative statement is defined as an evaluative statement indicating that something is bad about the other person or the content under discussion. (Negative evaluations of others with whom the person is interacting are called “faults.”) A deficiency of negative statements exists when the interactant fails to make negative statements that would be appropriate to the content of the discussion. An excess of negative statements exists when they are either too frequent, too lengthy, or inappropriate to the content under discussion.

3. *Opinions Given.* An opinion is defined as an interactant’s personal preference. A deficiency of opinions given is said to exist when an interactant expresses too few, or does not express opinions when requested to do so by the other partner. An excess of opinions exists when an interactant expresses personal preferences when they are uncalled for by the content of the discussion, when they are too frequent, or when they are irrelevant.

4. *Information Given.* Giving information is defined as making knowledge or facts available to the other interactant. A deficit is said to exist when an interactant withholds relevant information or refuses to give it when the partner requests it. An excess is said to exist when an interactant gives more

information than is necessary, either too frequently, in excessive detail, or when the information is irrelevant to the discussion.

5. *Opinions Requested.* Opinion requests are defined as statements or questions that request the partner to state his/her personal preferences. A deficit exists when one interactant appears not to know the relevant personal preferences of the other and makes no attempt to find out what these are. An excess exists when an interactant requests opinions from the other partner that, he/she already knows, when they are simply too frequent in occurrence, or when they are irrelevant to the discussion.

6. *Information Requests.* Information requests are defined as statements or questions by one interactant designed to elicit knowledge or facts from the other. These are deficient when the interactant fails to request information that he/she does not have and that is relevant to the topic. They are excessive when the requests are for information that the interactant already has, when they are too frequent in number, or when they are irrelevant to the topic under discussion.

7. *Acknowledgments.* Acknowledgment is defined as recognition of the other's point of view or statement. They are deficient when an interactant fails to recognize the other's point of view or give credit to the other's correct statements. They are excessive when the interactant gives credit too frequently or irrelevantly recognizes the other's point of view.

8. *Agreements.* Deficient agreements exist when an interactant fails to verbally express agreements with the statements of others. An excess of agreements is said to exist when an interactant agrees too frequently and/or irrelevantly with the statements of others.

9. *Disagreements.* Deficient number of disagreements exists when an interactant fails to verbalize his/her disagreement with the statements of others. An excessive number of disagreements exists when an interactant disagrees irrelevantly and/or frequently with the statements of others.

10. *Specific Talk.* A deficiency of specific talk is said to exist when an interactant fails to speak concretely and specifically in regard to a referent and, instead, he/she tends to be general and abstract. An excess of specific talk is said to exist when an interactant speaks too concretely and/or in too great a detail when more general statements are called for.

11. *Hypothetical Statements.* Hypothetical statements are defined as statements that pose possible situations or issues under consideration. They are typified by the "what if . . ." statements. They are deficient when an interactant fails to take into account the possible consequences of an action under consideration. They are excessive when an interactant continually or irrelevantly poses hypothetical situations.

12. *Conversational Control Statements.* Conversational control statements are defined as verbal cues such as questions, requests, directives, commands, and suggestions. They are deficient when one interactant fails to cue the other as to how he/she thinks the discussion should proceed. They are excessive when an

interactant attempts to control the form of the discussion by too frequently cueing the other.

13. *Content Focusing*. Appropriate content focusing is defined as adequately dealing with one topic or subtopic before moving to another. (In some cases, this can also include deciding not to deal with a particular issue.) The extremes are defined as shifting — switching topics or subtopics before the content has been adequately dealt with, or before the other has had an opportunity to discuss it; and persistence — sticking to or continually returning to a topic or subtopic after it has been adequately discussed.

14. *Amount of Talk*. Amount of talk refers to the proportion of time one interactant speaks relative to the other and to the overall time, considering the discussion as a whole. An interactant is said to overtalk when he/she talks more than the partner and takes up more of the total time than he/she is entitled to. It is possible for both interactants to undertalk, i.e., a discussion characterized by an excessive amount of silence. It is unlikely that both could overtalk except in the case where there is a great deal of simultaneous talk.

15. *Rate of Speech*. Rate of speech refers to the speed at which an interactant talks, e.g., words per second. In the extreme, slow talk could be so slow that it is difficult for the other to remember and process what is being said. Extremely fast talk would be so fast as to be partially unintelligible.

16. *Loudness of Speech*. Interactants can speak too loudly or too quietly to carry on a reasonable discussion, given the distance between them, size of the room, etc.

17. *Latency of Responses*. Latency refers to the time that elapses between the partner's completion of a speech unit and the interactant's next speech unit. In the extreme, slow latency is characterized by an inordinately long pause before the response. Extremely fast latency is said to occur when an interactant responds so quickly that he/she appears to clip the end of the partner's speech, making it unlikely that enough time has elapsed to adequately process the previous statement.

18. *Responsiveness*. Responsiveness applies to the length of speeches of an interactant in response to what is requested by others. An interactant is over-responsive when he/she speaks too long or goes beyond what is called for in responding to partner's talk. Underresponsiveness exists when an interactant says too little in relation to what a previous question or comment appears to call for.

19. *Pitch*. Pitch refers to the variation in pitch of an interactant's speeches. It can vary from monotone — practically no voice modulation — to sing-song, and in either extreme may distract the listeners from the semantic content of the statement.

20. *Affective Intonation*. Affective intonation refers to the match between a speaker's tone of voice and the semantic content of his/her speech unit. Underaffective intonation exists when the semantic content is such that it appears

to be emotion laden and the speaker's tone of voice does not reflect this. A speaker is overaffective when the semantic content does not appear to call for the affective tone of voice being used.

21. *Temporal Remoteness*. Temporal remoteness is defined as dwelling on irrelevant referents pertaining to the past or future. However, referring to events in the past that provide useful information for the topic under discussion is not considered temporal remoteness. Likewise, concrete plans, promises, etc., for the future are not temporally remote.

22. *Illogical Talk*. Illogical talk exists when an interactant makes an illogical statement considering what he/she or others have said.

23. *Quibbling*. Quibbling exists when an interactant attempts to explicate, clarify, or dispute a minor, tangential, and irrelevant detail.

24. *Off Topic Talk*. Off topic talk occurs when an interactant addresses him/herself to issues which are unrelated to the topic which has been agreed upon. The actual topic may have been either implicitly or explicitly agreed upon by the interactants. When the interactants explicitly agree to shift to a new topic or subtopic, it is not considered off topic talk.

25. *Dysfluent Talk*. Dysfluent talk exists when an interactant displays an excess of nonfluencies, such as stuttering, hesitating, or interspersing his/her speech with nonverbal vocalizations.

26. *Semantic Irrelevance*. Semantic irrelevance occurs when an interactant talks on a subject that does not show a clear semantic connection to the immediate focus of the discussion, e.g., irrelevant examples, ideas, or side comments.

27. *Redundant Talk*. An interactant is redundant when he/she repeats information, opinions, or preferences already given or known.

28. *Aversive Tone of Voice*. An interactant is said to have an aversive tone of voice when his/her voice quality is such that it is annoyingly unpleasant, e.g., threatening tone, whining.

29. *Obtrusion*. An obtrusion is an utterance made while another is speaking. Obtrusions become interruptions if they produce an immediate and apparently premature termination of the speech of the other.

30. *Overgeneralization*. An interactant is overgeneralizing when he/she misrepresents real world referents (behaviors or other events) by exaggerating such characteristics as their amount, importance, or quality. Overgeneralizations are typified by the use of such words as "always" and "never."

31. *Dogmatic Statement*. An interactant is dogmatic when he/she makes a statement in a categorical, unqualified, all or none, "black or white" manner.

32. *Pedantic Statement*. Pedantic statements are those which use big words where simpler, better-known words would be adequate, e.g., trepidation for fear, vicissitude for change.

33. *Ambiguous Statement*. Ambiguous statements are those which are equivocal, obscure, vague, or otherwise unclear in meaning and/or intent, e.g.,

“one thing wrong with our relationship is that you don’t have any problems.” (This could mean, “you don’t tell me your problems,” “you’re too good for me,” “you never admit you have problems,” etc.)

34. *Misrepresentation of Fact.* An interactant is misrepresenting a fact when he/she incorrectly represents real world events or the information previously given by other interactants: Incorrectly paraphrasing another’s statements of information is one subtype here.

35. *Misrepresentation of Evaluation.* An interactant is misrepresenting an evaluation when he/she incorrectly represents the evaluations, preferences, or opinions of other interactants. Paraphrase distortion of evaluation is one subtype here.

36. *Misrepresentation: Presumptive Attribution.* Presumptive attribution exists when an interactant misrepresents the meaning, motivations, feelings, and thoughts of others by incorrectly attributing nonobvious characteristics to them, e.g., “mind-reading,” “second guessing.” Correct assumptions about another’s feelings, thoughts, or other internal states are not considered presumptive attributions.

37. *Misrepresentation: Incorrect Autoclitic.* Incorrect autoclitics exist when a speaker makes incorrect remarks about what he/she is about to say. For example, a speaker might say, “I am not going to discuss my mother anymore,” and then go on to say several more things about her.

RELIABILITY AND VALIDITY

Reliability was examined in terms of interrater agreement, and concurrent validity of the ratings was assessed in terms of comparisons with precisely measured and carefully coded data from the same discussions. A major purpose of this investigation was to determine whether MCRaS could be used by clinicians to produce information which was similar to that produced by more rigorous coding procedures. If the ratings with MCRaS can be made reliably and can be shown to produce results similar to those obtained by more laborious coding procedures, such ratings can be justified for use in clinical assessments on the basis of their efficiency.

Method

The discussions that provide the data for the current study were all conducted as part of a series of decision-making training experiments with married couples. The original study was concerned with the evaluation of a coaching procedure for marital decision-making with clients referred to the research project from other agencies because of difficulties in decision-making regarding real-life issues. See O’Flaherty (1974), Thomas (1977), and Thomas *et al.*

(1976) for descriptions of the coaching procedure for decision making. The standard procedure was for each couple to have two assessment discussions ranging from 20–30 min in length before any decision-making training was undertaken, and a final assessment discussion after the completion of training, four to six weeks later.

The clients were interviewed in detail regarding each of their decision issues before any discussion took place. For all the topics discussed, each person agreed that the issue needed to be handled, that they preferred to handle it together, and that they each would allow the other to have a say in the decision. After the initial interviewing, introduction to the procedure, and selection of the topic to be discussed, the therapist left the room. Discussions were of real decision issues in the couple's life, conducted as a response display of decision-making verbal behavior. All discussions took place with only the two interactants present. They were seated opposite one another with a microphone at one end of the table; the experimenter/therapist was in the next room. Each session was tape recorded for later analysis.

Subjects

The subjects in this study were eleven married couples whose ages ranged from the mid-twenties to the mid-fifties. All were referred by social workers and psychologists to the research project. Each couple had several recurring real-life decision-making issues they were prepared to discuss, such as finances and budgeting, children's limits and behavior, social activities, work, or education. Ranging from two to four per couple, there was a total of 35 assessment discussions available for these analyses.

MCRaS Rating Procedure

Raters for this study were three final semester M. S. W. students, who were trained for approximately 20 hr over a five-week period. After initial familiarization with the checklist categories and definitions, raters listened to short excerpts of discussions and then identified and discussed specific examples of various categories. The checklist itself was learned in stages, starting with categories at the beginning of the checklist and, when high levels of agreement were reached, additional categories were added until raters were rating the entire checklist. Since many of the training tapes had been used in previous studies with the VPC, ratings in the training sessions could be checked against prior ratings as well as against those of fellow raters and the trainer.

After all discussion periods had been completed, the three raters listened to tape recordings of the discussions not used in training and filled out the MCRaS independently for each period. All raters completed the schedules simultaneously for each discussion, taking from 5 to 10 min each to complete.

Procedure for Reliability

After each of the three raters had rated the 35 discussions for husbands and wives, percentages of agreement were calculated for all possible pairs of raters on each of the 37 MCRaS categories. Both the use of three raters and the fact that agreements were not calculated until all of the discussions were rated were safeguards instituted to insure that raters did not drift into making similar judgments by influencing one another. Two indices of agreement were calculated, one for exact agreements and a second for approximate agreement. Exact agreement (no difference in scale ratings) is a very strict criterion, because a difference of one scale point, say +3 versus +2, would not ordinarily represent a meaningful disagreement in determining whether or not a particular category constitutes a problem for an individual. Therefore, the second set of percentages representing approximate agreement was calculated, which included as an agreement any instance in which the difference between the two raters was not more than one scale point. A discrepancy between raters of two or more scale points was thus considered to be an indication of meaningful disagreement for this second analysis of reliability.

Procedure for Validity

Categories chosen for the analysis of concurrent validity were selected to achieve a broad range of MCRaS response classes. One category from each of the four types of responses on the MCRaS was selected for postsession coding and comparison with the ratings. The MCRaS categories and their associated response groups are: (1) amount of talk (vocal characteristic); (2) negative statements (content); (3) overgeneralizations (referent representation); and (4) opinions requested (conversational control).

Coders who were not involved with the MCRaS ratings listened to tapes and coded the data for each comparison. Each validity analysis was done separately, and coding was always done for only one response category at a time. The means of the three MCRaS ratings were used as the data for comparison in these analyses.

Amount of Talk

The amount of time each individual talked was timed with an electronic timer for each of the 35 tapes. Each partner's talk time was timed separately and at least twice until a difference of less than .20 of a minute between two independent time measurements was reached. The resulting talk time for each individual was then divided by the combined talk time for both partners to establish the proportion of talk time for each. A second index representing the

proportion of total session time that each individual talked was computed by dividing the individual's measured talk time by the total amount of time elapsed during the discussion.

The proportions of talk time and of total session time were then compared to the mean MCRaS ratings for each subject. Analyses of variance were performed with each of the seven possible MCRaS scale ratings representing one cell.

Negative Statements

There were extremely high frequencies of negative talk during some of the discussions, making discrete coding of individual negative statements difficult. This necessitated using a 10-s interval coding procedure for occurrences of negative talk.

Two coders coded independently but simultaneously, seated at a table with a large stopwatch between them. At the end of each 10-s interval, each coder put a check mark in the box marked for that interval for each husband and wife if any portion of a negative statement occurred in that interval. Reliability was measured by totaling the number of agreed upon intervals and dividing by the total number of 10-s intervals in the discussion. The reliability measures between the two coders for each person in each discussion for the number of 10-s intervals in which negative statements occurred were calculated in terms of percentage of agreement. The mean percentage of agreement over the 35 discussions was 90.7. After all coding was completed and reliability between the coders had been established, one coder's data were randomly selected for each discussion to be used in comparison with MCRaS ratings of negative statements.

Two indices of negative talk for each subject were constructed for comparison with the ratings. An individual negative rate was established by dividing the total number of 10-s intervals during which a negative statement was coded by the total amount of talk time for that subject. The second was total negative rate, constructed by dividing the total number of 10-s intervals with a negative code by the total elapsed time for that discussion period. Analyses of variance were performed for both indices, using each of the seven scale points on the MCRaS as groups.

Overgeneralizations

Overgeneralizations were relatively infrequent and easily identified, and hence responses for this category were coded in 1-min blocks for each participant in a discussion. One coder coded all session tapes for overgeneralization, and 15 of the 35 were coded by a second coder to determine reliability. Reliability for

the two coders was 99.8% for the 15 tapes checked. The mean number of overgeneralizations for all 70 cases was 2.5. Twenty-six of the 70 cases had no overgeneralizations coded; and the mean number for the 44 cases that had at least one was 3.9.

Two indices of the rate of overgeneralizations were constructed in a manner similar to those for negative statements. The first, individual overgeneralization rate, was constructed by dividing the total number of overgeneralizations by the individual's own talk time. The second, total overgeneralization rate, consisted of the number of overgeneralizations divided by the total period time.

Opinion Requests

Although opinion requests were initially thought to be relatively easy to quantify, a problem arose when coders began to code opinion requests in the same manner that overgeneralizations had been coded. They had difficulty differentiating between instances of opinion and information requests, despite the fact that both rating categories on the MCRaS had fairly high reliabilities for all three pairs of raters. It was consequently decided to code all requests so that later they could be compared to the two categories on the MCRaS, both individually and in combined form. Three coders were used for this coding; one coded the 35 tapes, and the other two were used as check coders for reliability. All data used were from Coder 1. Coding was done in 1 min blocks. A 1 min block was considered to be an agreement if both coders had exactly the same number of requests in it.

Reliability of number of requests per 1-min blocks produced high percentages of agreement. The mean agreement over all 70 cases was 94.76%. The lowest intercoder reliability was 71%, and 30 of the 70 cases had reliabilities of 100%.

Rate of requests was rendered as two indices, analogous to those described for the other validity criteria. These were individual rate, the number of requests divided by the individual's talk time, and total request rate, the number of requests divided by total session time.

The vast majority of ratings fell in the 0 category (54 of 68 in opinion requests and 53 of 70 for information requests), with all the rest in category -1, indicating a truncated range with the raters judging the requests made essentially as neither excessive nor deficient. The actual coded requests, in contrast, ranged from 0 to 31, with a mean of 8.37 per individual.

Results

Reliability

The mean percentages of exact agreement over all categories were 58.31, 59.49, and 65.83 for rater pairs 1 and 2, 1 and 3, and 2 and 3, respectively.

Mean percentages of approximate agreement (i.e., within one scale point) for each pair of raters were 89.0 for raters 1 and 2, 89.53 for raters 1 and 3, and 92.72 for raters 2 and 3. Percentages of exact agreement and approximate agreement for each of the 37 categories are presented in Table I. Inspection of the data indicates that even for those categories in which the percentage of exact agreement is lowest, the percentages of approximate agreement are consistently high, i.e., 93% are above 80 and none is below 72.

Table I. Item by Item Percentage of Agreement for Each Pair of Raters for Exact Agreement and Agreement Within One Point on the Scale

MCRaS item	Rater Pairs					
	1-2		1-3		2-3	
	Exact	Within 1 pt.	Exact	Within 1 pt.	Exact	Within 1 pt.
1	44.3	77.1	42.9	78.6	41.4	87.1
2	45.7	72.8	45.7	90.0	61.4	94.2
3	60.0	81.4	67.1	81.4	68.6	95.7
4	54.3	78.6	60.0	84.3	62.9	87.2
5	61.4	85.7	68.6	87.1	81.4	98.6
6	57.1	88.5	57.1	90.0	78.6	95.7
7	35.7	80.0	47.1	84.3	65.7	91.4
8	54.3	78.6	60.0	85.7	68.6	86.6
9	58.6	90.0	54.3	88.6	68.6	92.9
10	41.4	81.4	55.7	85.7	52.9	85.8
11	32.9	84.3	42.9	87.2	42.9	90.1
12	40.0	80.0	47.1	80.0	51.4	88.6
13	34.3	80.0	47.1	82.8	47.1	90.0
14	35.7	87.1	57.1	91.4	50.0	91.4
15	27.1	81.4	18.6	80.0	61.4	95.7
16	44.3	95.8	35.7	94.3	61.4	97.1
17	32.9	81.5	22.9	78.6	44.3	85.7
18	37.1	79.9	21.4	72.8	34.3	84.3
19	35.7	85.7	35.7	88.6	67.1	95.7
20	72.9	92.9	62.9	92.9	60.0	90.0
21	45.7	88.6	52.9	88.6	55.7	74.2
22	95.7	100.0	90.0	100.0	88.6	100.0
23	44.3	95.7	61.4	95.7	70.0	98.6
24	48.6	94.3	57.1	85.7	51.4	88.5
25	82.9	97.2	80.0	98.5	88.6	100.0
26	98.6	100.0	92.9	100.0	91.4	100.0
27	48.6	92.9	50.0	90.0	60.0	90.0
28	74.3	95.7	75.7	91.4	74.3	91.4
29	54.3	94.3	51.4	88.6	51.4	90.0
30	67.1	95.7	52.9	94.3	57.1	92.8
31	72.9	90.1	64.3	90.0	68.6	91.5
32	98.6	98.6	95.7	98.5	97.1	100.0
33	91.4	100.0	95.7	100.0	90.0	100.0
34	92.9	100.0	87.1	100.0	88.6	100.0
35	87.1	95.7	82.9	94.3	87.1	100.0
36	55.7	91.4	67.1	92.8	52.9	91.4
37	92.9	100.0	94.3	100.0	92.9	100.0
Mean	58.31	89.00	59.49	89.53	65.83	92.72

Although the agreement percentages indicate moderate to high reliability between raters, it was necessary to determine whether the reliabilities were a result of some artifact of the scale itself. Thus if raters had a tendency to place their ratings around a particular scale point, artificially high reliabilities could have been obtained. Reliability that is artificially increased by a tendency of raters to cluster their ratings around scale points should be detected by finding a substantial negative relationship between the variance of a given item and the percentage of agreement between the raters on that item. To investigate this possibility, correlations were run between the percentages of agreement and the variances of the items, for both exact and approximate agreements. Separate correlations were run for items 1–20 (the bipolar items) and for items 21–37.

The correlations are given in Table II. For items 1–20, they were low for each pair of raters, and none approached significance. For items 21–37, the correlations were all negative and significant, indicating that for these items, some of the high reliability scores may have been produced by a tendency of raters to use one point on the scale.

Closer inspection of the data for items 21–37 indicated that infrequently used categories may have contributed to these results. There were 8 categories for which 95% or more of the ratings were 0 or 1, and, for these categories alone, the mean exact and approximate reliabilities were 92.5 and 98.9, respectively, for raters 1 and 2, 89.8 and 98.9 for raters 1 and 3, and 90.5 and 99.8 for raters 2 and 3. The eight categories that produced these results were: illogical talk (22), dysfluent talk (25), semantic irrelevance (26), pedantic statements (32), ambiguous statements (33), misrepresentations of fact (34), misrepresentation of evaluation (35), and incorrect autoclitics (37). Stated substantively, there is high agreement among raters on the infrequency of responses falling into these categories.

Table II. Correlations Between the Percentage of Agreement for Each Pair of Raters and the Variance of the Ratings on MCRaS Items

Rater pairs	Agreement	
	Exact	Approximate
Checklist items 1–20 ^a		
1-2	$r = -.354$	$r = +.088$
1-3	$r = -.119$	$r = +.156$
2-3	$r = -.217$	$r = +.054$
Checklist items 21–37 ^b		
1-2	$r = -.932$	$r = -.727$
1-3	$r = -.855$	$r = -.850$
2-3	$r = -.878$	$r = -.649$

^a $r = .444$ for $p \leq .05$.

^b $r = .482$ for $p \leq .05$.

Inspection of the remaining nine categories in the second half of the MCRaS provided additional information. The means for reliability were calculated using only these remaining categories. The means for exact agreement dropped to somewhat less than the overall means for all 37 items. For the three rater pairs, the percentages were 56.83, 59.2, and 60.16. The means for the percentage of approximate agreement remained high, however. They were 93.19, 90.79, and 89.82, for rater pairs 1 and 2, 1 and 3, and 2 and 3, respectively. Thus it appears that while the items for which ratings were infrequently made did inflate the overall reliability scores for exact agreement, scores of approximate agreement were not appreciably affected by the presence of such categories.

Validity

In regard to the amount of talk, the results of the analysis of variance using both indices of talk time are present in Table III. Both measures are highly related to the MCRaS rating ($p < .001$). The raters judgments were evidently clearly dependent upon how much of the time the individual actually talked.

The results of the analysis of variance for individual negative rate and total negative rate are presented in Table IV, where it may be seen that although both negative rates reach significance, the total negative rate ($F = 16.879, p < .001$) is clearly superior to individual negative rate ($F = 3.286, p < .004$). The means for total negative rate are in the predicted direction, except for the anomaly of a higher mean, based upon one rating, for the rating category of -1 . The means for the individual negative rate are in the predicted direction, except for the means at the scale extremes where the n 's are small – the mean for the single anomalous rating mentioned above and the somewhat lower mean of 3.8 for the five ratings falling into the $+3$ category.

Table III. Ratings of Amount of Talk Compared to Measured Proportion of Talk Time and Proportion of Tape Time

MCRaS rating	<i>N</i>	Mean proportion of talk time ^a	Mean proportion of total tape time ^b
-3	4	.170	.132
-2	8	.263	.176
-1	14	.399	.290
0	17	.483	.349
+1	14	.639	.472
+1	7	.659	.505
+3	6	.811	.611
Total	70	.500	.368

^a $F = 39.083, p < .0001.$

^b $F = 35.538, p < .0001.$

Table IV. Ratings of Negative Statements Compared to Individual Negative Rate and Total Negative Rate

MCRaS rating	<i>N</i>	Mean individual negative rate ^a	Mean total negative rate ^b
-3	0		
-2	0		
-1	1	2.727	.679
0	29	1.266	.411
+1	16	3.552	.870
+2	19	4.252	1.572
+3	5	3.867	2.264
Total	70	2.806	.967

^a $F = 4.2858, p < .004.$

^b $F = 16.879, p < .001.$

The results for overgeneralizations are presented in Table V. Means for all groups are in the predicted direction for both indices, and the results for both are significant. The comparison for the total overgeneralization rate is more significant than that for the individual overgeneralization rate ($F = 10.578, p < .001$ versus $F = 3.212, p < .05$).

Individual rate of requests and total request rate were compared to the MCRaS ratings of opinion requests and of information requests, as well as to a combined rating of opinion and information requests. The combined rating was constructed by averaging the two ratings for each person in each discussion. The means were largely in the predicted direction. However, none of the comparisons was statistically significant. These comparisons were opinion requests versus individual request rate ($F = 3.398, n.s.$), opinion request versus total request rate ($F = 2.555, n.s.$), information requests versus individual request rate ($F = 1.481, n.s.$), information requests versus total request rate ($F = 1.054, n.s.$), and combined requests versus total request rate ($F = 1.592, n.s.$).

Table V. Ratings of Overgeneralizations Compared to Individual Overgeneralization Rate and Total Overgeneralization Rate

MCRaS rating	<i>N</i>	Mean individual overgeneralization rate ^a	Mean total overgeneralization rate ^b
0	42	.215	.070
+1	20	.303	.113
+2	8	.534	.289
+3	0	—	—
Total	70	.534	.108

^a $F = 3.212, p < .05.$

^b $F = 10.578, p < .0001.$

Clinical Use

MCRaS can be used at any point in marital therapy when communication may be an area of concern. After the agreement to work on communication problems and the couple's description of what they see as their communication problems, one or more response display discussions should be held with the couple. In the response display, husband and wife discuss a topic of mutual concern, chosen in conjunction with the clinician, for approximately 15 to 20 min. The clinician may leave the couple alone to carry out the discussion and listen to the discussion later through replay of audiotape recording, or listen "on line" through head phones or in person. In the latter case, the practitioner should sit as far out of the line of vision of both interactants as possible.

During the discussion, the clinician listens to the verbal interchanges in terms of the response categories of MCRaS. Immediately following the discussion (or after listening to a tape recording), the clinician rates each of the interactants on all of the 37 categories. The postdiscussion rating typically takes about 5 min. The use of the checklist to cue the listener during the discussion as well as familiarity with the categories and definitions makes the task of rating easier than at first it might appear. The number of problems identified for a given individual that merit high ratings is typically between two and five, so that although the rater is cued to listen in terms of most or all of the 37 categories, responses that occur frequently can be discriminated and categories for which very few responses occur can generally be rated very easily. MCRaS thus permits a rapid narrowing of a potentially large number of verbal problems down to a few for each couple.

After the discussion has taken place, the couple should be interviewed to determine whether or not the discussion was typical for them. If the partners agree that the discussion was similar to those they frequently have, two response display discussions are likely to be sufficient to determine target areas for specification and intervention. If obtaining response display discussions in the office setting is difficult, either because the setting itself produces changes in the couple's usual interaction or because of limited session time, the couple can be asked to record a discussion at home which the clinician can rate later using MCRaS.

The verbal problems identified by means of MCRaS are candidates for target areas of intervention. Deficits, such as lack of positive talk, ordinarily do not require further specification. However, surfeits generally merit individualized examination for each couple. Consider negative talk, for example, although an excessive amount of negative talk is one of the categories most commonly rated as problematic for married couples (Thomas, *et al.*, 1974), the particular content and pattern may be distinctive for each couple. For one couple in this series, the husband received a rating of +3 for negative talk. Further specification determined that his negative talk consisted almost entirely

of negative self-statements such as, "I always seem to mess things up," "I just don't do things, right," and "It's my fault." The wife, in contrast, was rated only +1 on this category, but she apparently reinforced the husband's negative statements either by agreeing with them or by asking questions such as, "Why do you think you do that?", which served to generate more negative self-statements from the husband. In this case, one direction for treatment might be to reduce the husband's negative talk by altering the wife's responses to his negative self-statements. In contrast, negative talk for another couple consisted of escalating exchanges of negative comments about one another. For example, Wife: "You don't pay enough attention to the kids;" Husband: "I don't see why you can't stop them from fighting;" Wife: "You don't even seem to like them;" Husband: "You're raising three brats;" Wife: "You're not even a good imitation of a father." In this case, one goal of communication training might be to increase the responsiveness of each person to the content of what the other says. Additionally, the content and pattern here suggest problems of child management and parent-child relationships as candidates for intervention. After the content and pattern of verbal responses have been thus examined, target behaviors can be identified and directions for intervention established, following more general guidelines to treatment of marital discord such as those discussed by Gottman *et al.* (1976), Jacobson and Margolin (1979), Weiss and Birchler (1978), and Thomas (1977).

MCRaS can be used to monitor progress during intervention, using periodically rated response displays, as described above. It can also be employed after intervention to determine the degree of generalization of change, from office to home, by having the couple record discussions to be rated. The ease of recording discussions for later rating also makes this useful in follow-up. Couples who have terminated treatment can be asked to send in tapes at specified intervals, which can be evaluated using MCRaS to determine whether changes have been maintained.

DISCUSSION

The results were largely very encouraging for the reliability and validity of this observational rating instrument. The interrater reliabilities for approximate agreement (within one scale point) were high, and are judged to be high enough for clinical purposes and, indeed, for many research inquiries. However, the reliabilities for exact agreement were moderate for many categories, and may be less than adequate for some research purposes. Because agreement within one scale step, as expressed by the measure of approximate agreement, is sufficient to indicate problem areas for specification and modification, it may be concluded that MCRaS can be employed clinically with satisfactory rater reliability.

The validity analyses for the categories of amount of talk, negative statements, and overgeneralizations indicated that the ratings on the MCRaS were

highly related to the measured data involving rates of specific responses employed as indicators of validity. The two indices of rate constructed from the coded data were used in an attempt to clarify further what variables controlled the behavior of the MCRaS raters. For the three rating categories in question, the superiority of the rate calculated over the total session time indicates that the raters based their judgments on response frequencies involving the discussion as a whole rather than on those frequencies that were in direct proportion to the amount of time the individual in question was talking.

The measure of validity involving the coded number of requests was only weakly and not significantly related to the ratings on the MCRaS for opinion requests or information requests. One reason for the failure of these ratings to be discriminative is the truncated scale range for these ratings. As indicated earlier, most fell into the 0 category on the rating scale (54 of 68 for opinion requests, and 53 of 70 for information requests), with essentially all of the rest falling into the -1 category. With such a clustering of ratings on the scale point indicating adequacy and so little range, it is difficult statistically to obtain a strong association with a criterion variable, even if its range and variation were much greater, as they were in the case of coded requests. A different cohort of marital partners in which the variations of these ratings were greater would be a better sample on which to test the validity of the ratings.

It is also possible that in the case of rating opinion and information requests, the ratings of deficiency or excess were not based directly on sheer frequency per unit of time, as they apparently were for the other rating categories, but rather were guided by the rater's judgment of whether there had been enough opinion (or information) shared and, in this light, whether the requests were relatively adequate. If judgments were made here on such a relative basis, then the rate of requests alone is not an appropriate validation criterion. Failure to validate using a measure of quantity alone, such as the frequency of an observed event, does not preclude the possibility that the rating is valid by a relative criterion. In the case of the ratings of opinion requests and information requests, the ratings may be made validly using validation criteria more appropriate for evaluating requests relative to some standard. Although validation of judgments of interaction using relative criteria poses vexing methodological problems at this stage of research sophistication, the distinction between "absolute" and "relative" ratings merits further attention in regard to its implications for the construction and validation of rating instruments.

The results of this study largely provide further support, beyond those reported earlier for the predecessor instrument (Thomas *et al.*, 1974), for the reliability and validity of the MCRaS. The schedule may be completed in 5-10 min immediately after a discussion by the marital partners and hence may be a useful tool for purposes of clinical assessment, monitoring of treatment outcomes, and clinical research. However, it should be pointed out that this instrument is still in its formative stages, and further research needs to be conducted

to validate additional categories of the MCRaS and to examine further the validity of observationally based clinical rating instruments. The couples who served as subjects for this study were all identified as having verbal communication problems. While this was a particularly appropriate group for the development of this instrument, the utility of MCRaS would be enhanced by future research focused on the collection of normative data involving nondistressed as well as distressed couples.

After reviewing current research on observational coding systems of marital interaction, Jacobson *et al.* (1981) concluded that "while the direct observation of marital interaction will remain an important area for future research investigations, for the present it is unlikely to be particularly useful to practicing clinicians. The legacy for clinicians is that communication and problem solving should be evaluated as carefully as possible, despite the implausibility of producing the kind of data depicted in research investigations." The MCRaS is a clinical rating instrument that may help at this point, inasmuch as it can be used reliably to help clinicians discriminate among many potential areas of problematic verbal communication and to pinpoint those which are most severe for a given couple. It does not, of course, eliminate the need for detailed specification of identified problems prior to intervention.

REFERENCES

- Birchler, G. R. Communication skills for married couples. In Bellack, A. S., & Hersen, R. M. (Eds.), *Research and practice in social skills training*. New York: Plenum, 1979.
- Carter, R. D., & Thomas, E. J. A case application of a signaling system (SAM) to the assessment and modification of selected problems of marital communication. *Behavior Therapy*, 1973, 4, 629-645.
- Eisler, R. M., Hersen, M., & Agras, W. S. Audiotape: a method for the controlled observation of nonverbal interpersonal behavior. *Behavior Therapy*, 1973, 4, 420-425.
- Gottman, J., Notarius, C., Gonso, J., & Markman, H. *A couple's guide to communication*. Champaign, Ill.: Research Press, 1976.
- Hops, H., Wills, T. A., Patterson, G. R., & Weiss, R. L. Marital interaction coding system. Unpublished manuscript, University of Oregon and Oregon Research Institute, 1972.
- Jacobson, N. S., Elwood, R., & Dallas, M. The behavioral assessment of marital dysfunction. In Barlow, D. H. (Ed.), *Behavioral assessment of adult disorders*. New York: Guilford Press, 1981.
- Jacobson, N. S., & Margolin, G. *Marital therapy: strategies based in social learning and behavior exchange principles*. New York: Brunner/Mazel, 1979.
- Jacobson, N. S., & Martin, B. Behavioral marriage therapy: current status. *Psychological Bulletin*, 1976, 83, 540-577.
- O'Flaherty, K. W. Evaluation of a coaching procedure for marital decision making. Unpublished dissertation, University of Michigan, 1974.
- Olson, D. H., & Straus, M. A. A diagnostic tool for marital and family therapy: the SIMFAM technique. *Family Coordinator*, 1972, 21(3), 251-258.
- Strodbeck, F. L. Husband-wife interaction over revealed differences. *American Sociological Review*, 1951, 16, 468-473.
- Thomas, E. J., Walter, C. L., & O'Flaherty, K. A verbal problem checklist for use in assessing family verbal behavior. *Behavior Therapy*, 1974, 5, 235-246.

- Thomas, E. J., O'Flaherty, K., & Borkin, J. Coaching marital partners in family decision making. In Krumboltz, J. D., & Thoresen, C. W. (Eds.), *Counseling methods*. New York: Holt, Rinehart and Winston, 1976.
- Thomas, E. J. *Marital communication and decision making*. New York: The Free Press, 1977.
- Weiss, R. L., & Bichler, G. R. Adults with marital dysfunction. In Hersen, M., & Bellack, A. S. (Eds.), *Behavior therapy in the psychiatric setting*. Baltimore: Williams & Wilkins, 1978.
- Weiss, R. L., & Margolin, G. Marital conflict and accord. In Ciminero, A. R., Calhoun, K. S., & Adams, H. E. (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.