THE HELPFUL RESPONSES QUESTIONNAIRE: A PROCEDURE FOR MEASURING THERAPEUTIC EMPATHY

WILLIAM R. MILLER, KIM E. HEDRICK AND DEBRA R. ORLOFSKY

The University of New Mexico

The development and evaluation of a measure of accurate empathy is described. The Helpful Responses Questionnaire (HRQ) is a brief free response questionnaire that can be administered to groups. Interrater reliability coefficients for items range from .71 to .91, and reliability for total HRQ scores was found to be .93. HRQ scores rose significantly following a 2-day workshop on active listening and crisis intervention skills offered in 14 communities. Normative data are reported based on a sample of 190 paraprofessional trainees. Level of empathy was found to be modestly related to respondents' self-esteem.

The construct of accurate empathy (Rogers, 1957) has been the subject of extensive research that generally has supported the importance of this therapeutic skill in producing beneficial client outcomes (Garfield & Bergin, 1986; Hart & Tomlinson, 1970). More recent research has demonstrated the impact of accurate empathy for counselors who work within frameworks other than client-centered therapy (Miller, Taylor, & West, 1980; Valle, 1981). The key concept of "active listening" in Gordon's (1970) Parent Effectiveness Training was derived directly from the construct of accurate empathy.

The most frequently used approach for measuring empathy has been through observer ratings. Truax introduced a reliable and widely used rating scale for coding individual counseling sessions (Truax & Carkhuff, 1967). Reliable and valid paper-and-pencil measures of accurate empathy or active listening have been more elusive. Several instruments are available that provide simulated client statements and ask the respondent to indicate which of several alternatives is the "correct" meaning of the statement (e.g., Jones & Mohr, 1976). Such instruments do not measure the ability to *generate* empathic responses, but only an ability to recognize latent content in communication. The present approach was designed as an open-response questionnaire for individual or group administration, analogous to the Truax scale, and conceptually linked to Gordon's (1970) description of active listening. It was developed as part of a project to develop, implement, and evaluate training materials for crisis intervention counselors in rural community settings.

Scale Description

The Helpful Responses Questionnaire (HRQ) consists of six paragraphs that simulate communications from individuals with specific concerns. After each paragraph, a space is provided for the respondent to write a helping response. Instructions for the questionnaire are:

> The following six paragraphs are things that a person might say to you. For each paragraph imagine that someone you know is talking to you and explaining a problem that he or she is having. You want to help by saying the right thing. Think about each paragraph as if you were really in the situation, with that person talking to you. In each case write the *next thing* that you would say if you wanted to be helpful. Write only one or two sentences for each situation. Please print or write clearly.

The six stimulus paragraphs then are provided, each followed by the instruction, 'Write here what you would say next." The six communications are listed in Table 1. Average administration time is 15 to 20 minutes.

Table 1
Items of the Helpful Responses Questionnaire

1. A 41-year-old woman says to you: "Last night Joe got really drunk and he came home late and we had a big fight. He yelled at me and I yelled back and then he hit me really hard! He broke a window and the TV set, too! It was like he was crazy. I just don't know what to do!"

2. A 36-year-old man tells you: "My neighbor is really a pain. He's always over here bothering us or borrowing things that he never returns. Sometimes he calls us late at night after we've gone to bed and I really feel like telling him to get lost."

3. A 15-year-old girl tells you: "I'm really mixed up. A lot of my friends, they stay out real late and do things their parents don't know about. They always want me to come along and I don't want them to think I'm weird or something, but I don't know what would happen if I went along either."

4. A 35-year-old parent says: "My Maria is a good girl. She's never been in trouble, but I worry about her. Lately she wants to stay out later and later and sometimes I don't know where she is. She just had her ears pierced without asking me! And some of the friends she brings home—well I've told her again and again to stay away from that kind. They're no good for her, but she won't listen."

5. A 43-year-old man says: "I really feel awful. Last night I got drunk again and I don't even remember what I did. This morning I found out that the screen of the television is busted and I think I probably did it, but my wife isn't talking to me. I don't think I'm an alcoholic, you know, because I can go for weeks without drinking. But this has got to change."

6. A 59-year-old unemployed teacher tells you: "My life just doesn't seem worth living any more. I'm a lousy father. I can't get a job. Nothing good ever happens to me. Everything I try to do turns rotten. Sometimes I wonder whether it's worth it."

The HRQ is scored by rating each response on a 5-point ordinal scale of depth of reflection. Scale definitions integrate Truax's depth rating system with concepts from Gordon (1970). A rating of 1 is assigned when the response contains no reflection, but does include at least one element scorable as a "roadblock" response as defined in Gordon's (1970) "typical twelve" responses. (Cf. Miller & Jackson, 1985.) This level of response actively interrupts the flow of communication. A score of 2 is assigned when the response contains both a reflection (scorable at level 3, 4, or 5 below) and a roadblock, or contains neither reflection nor roadblock response. A response that is a reflection or contains a reflection that merely repeats the content already stated is scored 3. When the reflection reaches paraphrase status, adding inferred meaning that appears appropriate or plausible, a score of 4 is assigned. Finally, a 5 is scored when the response qualifies at level 4 and also includes either a reflection of feeling that fits the original statement or an appropriate metaphor or simile. When multiple responses are made, the highest level of reflection is scored unless a roadblock is included (resulting in a score of 2).

Responses at level 1 or 2 can be scored further to indicate which of Gordon's 12 roadblocks they contain (Miller & Jackson, 1985). This can be useful as training feedback.

Method

Subjects and Procedure

Workshops designed to train paraprofessional crisis intervention counselors were offered in 14 rural New Mexico communities. The workshops were advertised through news media, posters, and personal contact visits to each community and were open to all interested individuals. Content of the workshops included 6 to 8 hours of training in active listening skills and 6 to 8 hours of instruction and practice in other crisis intervention skills, distributed over 2 days. The two junior authors served as trainers.

The HRQ was administered at the beginning of the first workshop day before any training had been initiated and at the end of the second day after training had been

completed. Two other instruments were administered only before training: the Profile of Mood States (McNair, Lorr, & Droppelman, 1971) and the Self-Esteem Inventory (Coopersmith, 1975).

All 1,860 responses were rated independently by two research assistants who were not involved in training. One trainer (KEH) also rated 120 responses as a further check on reliability. Prior to coding, all responses were typed onto individual index cards so that raters could not identify which responses had been written by the same respondent, nor could they discriminate which responses had been given before vs. after training. This blind rating system was employed to remove possible biases due to knowledge of training status or to halo effects caused by scoring multiple responses known to be made by the same individual.

Results

A total of 190 individuals (153 women) completed the pre-training questionnaires. Their average age was 37.3 years, and they reported a mean of 14.4 years of education. Ethnicity of trainees was as follows: 109 White non-Hispanic, 56 Hispanic, 16 Native American, and 9 others. Of the 190 who began training, 120 completed the HRQ both before and after the workshop. This provided a total of 310 protocols, which contained 1,860 helping responses.

Interrater reliability was estimated by computing product-moment correlations for all responses scored by pairs of raters. Reliability coefficients for individual items ranged from .71 to .91 (all $p < .001$). The reliability of the principal raters when checked against a trainer on 120 randomly chosen responses proved to be .85 and .83 ($p < .001$). The interrater reliability for total HRQ scores (sum of the six item scores for each respondent) was .932 ($p < .001$).

The mean scores of respondents on the six scale items before and after training are reported in Table 2, along with reliability coefficients for each item. To provide a larger normative sample, data from the full population are reported at pre-training and did not differ significantly from pre-training means of the smaller group who also completed post-training questionnaires ($n = 120$). For the latter group only, pre-training scores were compared with those from post-training to determine whether there had been a significant improvement in empathic responses. A multivariate analysis of variance was employed to combine scores from the two raters. A highly significant main effect of training was found, $F(1, 118) = 101.2$, $p < .001$, an indication that training produced substantial improvement as reflected on the HRQ.

Internal consistency of the HRQ was found to be satisfactory. The mean inter-item correlation was .67 at pre-training and .57 at post-training. Cronbach's alpha statistic,

Table 2
Mean Empathy Item Scores Before and After Training

	Before training ($n = 190$)			After training ($n = 120$)		
HRQ item	Rater 1	Rater 2	*r*	Rater 1	Rater 2	*r*
1	1.56	1.65	.89	3.03	3.06	.83
2	1.53	1.55	.84	3.05	3.19	.88
3	1.56	1.68	.76	2.78	3.37	.71
4	1.52	1.73	.78	3.09	3.35	.84
5	1.61	1.72	.91	2.98	3.08	.88
6	1.78	1.92	.88	2.98	3.11	.90

the mean of all possible split-half reliability coefficients, was .92 at pre- and .89 at post-training. Mean correlations between item scores and total scale score (corrected by removing the item being correlated) were found to be .87 and .79, respectively.

Test-retest reliability was evaluated by correlating trainees' two scores, recognizing that a period of training intervened between first and second testing. A correlation coefficient of .45 was obtained. Combined with the significant training effect presented above, this modest coefficient suggests differential skill acquisition such that some individuals showed improvement in empathy, whereas others did not. In an attempt to account for some of this remaining variance, we computed product-moment correlations to determine whether relationships existed between degree of accurate empathy (either before or after training) and scores on the mood and self-esteem measures. At pre-training, only one correlation coefficient of seven exceeded an α level of $p < .01$: Self-esteem was found to be related positively to empathy scores, $r\,(190) = .19, p < .01$. This relationship remained at post-training, $r\,(120) = .20, p < .01$, and, in addition, the Confusion subscale of the Profile of Mood States showed an inverse relationship with empathy, $r\,(120) = -.20, p < .01$. Despite statistical significance, these relationships are very small, and each accounts for only 4% of variance in empathy.

Discussion

Direct observational assessment of empathy (Truax & Carkhuff, 1967) remains a desirable approach whenever possible. The HRQ is an alternative when individual observation is not feasible, as when groups are being assessed and trained. It avoids the simplification and matching-to-sample rigidity of tasks that require identification of the "correct" meaning of a stimulus statement, by rating freely generated responses in a manner similar to that used for observing sessions. Reliability coefficients both for items and for total HRQ scores proved quite acceptable, as were measures of internal consistency. Scores on the HRQ increased after training in active listening even though questionnaires were scored by raters blind to the identity of respondents and to pre- vs. post-training responses. Pre-training levels of self-esteem and confusion were related modestly to empathy scores and accounted for a small portion of variance in empathy responses before and after training. Other unmeasured variables probably account for variance in empathic skills, as defined by the HRQ, and further study is needed to clarify correlates of trainees' initial level and acquisition of empathy. Finally, it must be noted that a change in questionnaire scores guarantees neither generalization of change to observational measures nor maintenance of change over time. Future studies could explore the convergence of questionnaire and observational data and the stability of training-related changes in therapeutic empathy.

References

Coopersmith, S. (1975). *Self Esteem Inventory.* Lafayette, CA: Self Esteem Institute.

Garfield, S. L., & Bergin, A. E. (1986). *Handbook of psychotherapy and behavior change* (3rd ed.). New York: John Wiley.

Gordon, T. (1970). *Parent effectiveness training.* New York: Wyden.

Hart, J. L., & Tomlinson, T. M. (eds.). (1970). *New directions in client-centered therapy.* New York: Houghton Mifflin.

Jones, J. E., & Mohr, L. (1976). *The Jones-Mohr Listening Test.* La Jolla, CA: University Associates.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual for the Profile of Mood States.* San Diego, CA: Educational & Industrial Testing Service.

Miller, W. R., & Jackson, K. A. (1985). Not listening. In *Practical psychology for pastors: Toward more effective counseling* (pp. 31-38). Englewood Cliffs, NJ: Prentice-Hall.

Miller, W. R., Taylor, C. A., & West, J. C. (1980). Focused versus broad-spectrum behavior therapy for problem drinkers. *Journal of Consulting and Clinical Psychology, 5,* 590-601.

ROGERS, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology, 21,* 95-103.

TRUAX, C. B., & CARKHUFF, R. R. (1967). *Toward effective counseling and psychotherapy: Training and practice.* Chicago: Aldine.

VALLE, S. K. (1981). Interpersonal functioning of alcoholism counselors and treatment outcome. *Journal of Studies on Alcohol, 42,* 783-790.

MEDICAL, FAMILY, AND SCHOLASTIC CONDITIONS IN URBAN DELINQUENTS

JOHN R. HUGHES

University of Illinois at Chicago College of Medicine

ROBERT ZAGAR

Juvenile Division of the Circuit Court of Cook County

ROBERT B. SYLVIES

Illinois School of Professional Psychology

JACK ARBIT

Northwestern University Medical School and Loyola University Stritch College of Medicine

KENNETH G. BUSCH

University of Illinois at Chicago College of Medicine and Juvenile Division of Circuit Court of Cook County

NORMAN D. BOWERS

Northwestern University

Effects of medical, family, and scholastic conditions were evaluated for the number and type of offenses and test score performances in 1,962 urban delinquents. Conditions were evaluated using multivariate followed by univariate analysis of variance with post-hoc tests. There were subgroup differences in the number and type of offenses and test score performances. Also, interactions of medical, family, and scholastic conditions changed the number and type of offenses and lowered test scores. Findings supported a developmental biopsychosocial model of delinquency that emphasized the predominance of medical/scholastic conditions. Orphaned or one-parent delinquents with nervous system or neonatal conditions, retardation, or hyperactivity committed assault.

The authors wish to express gratitude for the assistance of Judge Arthur Hamilton, Tom Jones, John Hahn, James Jordan and David Schwartz for providing access to the data; to Toya Woullard, Edward McClinton, Tom Baglajewski, Sherry Perry, Susan Flannagan, Darek Jakubowski, Lynn Hume, and Di Theodoros for data collection and computer entry; to Susan Webb-Kmiec for the typing and revisions of the manuscript; to Judge Julia Quinn Dempsey, Brooke Whitted, Paul Stephanidas, John Hamje, Drs. Robert E. Bussell, Eliezer Schwartz, and Jack Saporta, who commented on earlier drafts of this paper. This research is funded in part by the Edith Schiller Neurology Fund (Northwestern Medical Hospital) and by the Department of Psychiatry and Behavior Sciences, Northwestern University Medical School.

Requests for reprints should be addressed to Dr. Jack Arbit, Department of Psychiatry and Behavioral Sciences, Northwestern University Medical School, 303 East Chicago Avenue, Chicago, IL 60611.