Behavior Research Methods & Instrumentation
1981, Vol. 13 (4), 436-442

New approaches to the design of computerized interviewing and testing systems

ROBERT L. STOUT
Brown University, Providence, Rhode Island 02912, and
Butler Hospital, Providence, Rhode Island 02906

Most computer interviewing and testing systems have adopted paper-and-pencil approaches to information gathering with little modification. However, computer technology offers two fundamental advantages over paper-and-pencil technology for psychological information gathering: (1) A computer can record ancillary data such as latencies and pressure on response keys during an interviewing session, and (2) A computer can react adaptively to special events as these arise during a session. Ways to capitalize on these advantages are outlined. A pilot study of interviewee behavior during a computer problem-screening interview is described, and the implications of the results for future research in the area are discussed. Passive and active computer testing systems occupy positions on a continuum between paper-based psychological testing and the flexible, but less well controlled, technology represented by the human. With its unique capabilities, computer technology has a special role to play in the future of psychological measurement.

Inexpensive microprocessors are popularizing the spread of computer programs that gather clinical information by interacting directly with clients and/or clients' relatives or friends. Interactive computer information gathering systems have been found to be practical and advantageous in a variety of clinical applications, including computerized testing and computer interviewing.

Computer interviewing and computerized testing are not wholly distinct; the phrase "computerized testing" has most often been used when the purpose of the data gathering is to estimate a subject's score on one or more dimensions measured by a standardized instrument, such as the MMPI (see, e.g., the systems described by Greist & Klein, 1980; Johnson & Williams, 1978), whereas the phrase "computer interviewing" is more frequently used when the goal is to obtain a detailed listing of problematic behaviors or a behavioral inventory (see, e.g., Angle, Ellinwood, W. Hay, Johnsen, & L. Hay, 1977; W. Hay, L. Hay, Angle, & Ellinwood, 1977). The issues to be discussed in this paper are common to both computer interviewing and computer testing; for the sake of brevity, "interviewing" will be used below to mean both interviewing and testing.

The research reported in this article was supported in part by NIMH Grant MH 26012, "Problems as Predictors of Treatment and Outcome," Richard Longabaugh, principal investigator. The author would like to express his gratitude to Willa Kay Wiener-Ehrlich, who devised much of the interviewing software, and to Linda and William Hay, who devised the original version of the computer problem screening questionnaire. The author would also like to acknowledge the assistance of Lynelle Jenik in gathering the data, as well as that of Duane Bishop, Edward Fink, and Gabor Keitner, who recommended interesting patients for the study.

In most clinical information gathering systems, the interviewees interact with the computer primarily by responding to questions having multiple-choice response formats. There have been attempts to create computer systems capable of interviewing a person in the more traditional sense of conducting a natural language dialogue (Colby, 1980); however, these systems are currently only of research interest and will not be considered in this paper. Also, systems whose primary purpose is physiological monitoring are outside the purview of this discussion.

A number of investigators have demonstrated that interactive computer information gathering provides benefits of economy, speed, reliability, and even acceptability to interviewees over paper-and-pencil data gathering techniques (Greist & Klein, 1980; W. Hay et al. 1977). These benefits are by no means negligible, but they represent quantitative rather than qualitative gains over standard paper-and-pencil data gathering methods. In most current applications, the computer is used to gather the same kind of information one would gather on a paper form, and with a small number of exceptions the computer does not use the information it gathers interactively any differently from information coded and keypunched from a paper form. Thus, the primary role of the computer in most interactive clinical data gathering systems has been to perform routine bookkeeping and arithmetic. There is no doubt that the computer does do these tasks very well, but the fundamental promise of computer technology lies in the fact that it is capable of much more.

The purpose of this paper is to describe some novel ways in which computer technology might be applied in clinical data gathering applications and to discuss the advantages and disadvantages of these new approaches

0005-7878/81/040436-07$00.95/0

MAJOR LIMITATIONS OF CURRENT APPROACHES

A piece of paper records what is written on it, and no more. Under ideal circumstances, it is sufficient to know only what response the interviewee has marked for each question. In most clinical data gathering situations, however, the circumstances are less than ideal, and it would be useful to know about any complications that might affect the interpretation of the responses. When an interviewee sees a question presented on the screen of a video terminal or reads a question on a paper form, a variety of events may occur, including the following: (1) The interviewee may understand the question adequately and respond appropriately. (2) The interviewee may misinterpret the question and respond erroneously. (3) The interviewee may understand the item but be reluctant to fit his/her response into the categories provided. (4) The interviewee may disregard the question and respond arbitrarily. (5) An unusual emotional state or idiosyncratic train of associations in the interviewee may cause a biased or atypical response. (6) The interviewee for any number of reasons may respond evasively. (7) The interviewee may make a typing mistake. (8) The interviewee may refuse to answer the question.

Undoubtedly, these eight categories do not exhaust all the possibilities. Clearly, one's interpretation of a response would be affected if one knew that one of Events 2-7 above had happened when that response was made. Unfortunately, with paper-and-pencil technology, it is difficult to detect the presence of one of these invalidating events unless the error is gross.

Data invalidity is, of course, not a new problem, and several techniques have been developed to deal with the problem in the framework of paper-and-pencil technology, including redundancy, lie scales, social desirability scales, and screening. Computer interviewing systems developed to date have adopted these techniques with little if any modification. A screening instrument, the Q1, has been developed to detect interviewees who are overtly hostile to computer interviewing or who are likely to be unreliable informants (Johnson, Williams, Klingler, & Giannetti, 1977); also, when instruments such as the MMPI have been adapted for computer administration, the scales and consistency checks built into these instruments have been adopted without modification. These traditional measures to deal with data validity problems are useful to some extent, but they also have some major drawbacks. High scores on social desirability scales do not seem to be reliably indicative of a generalized tendency to respond falsely (Bradburn & Sudman, 1979, pp. 85-106). Other global consistency/validity measures may be more valid, but all suffer from some general limitations. One major drawback is that any global score computed after testing that implies that invalid responses may be present in the data is of use only for discarding data one has already invested time and energy to obtain; a global score is of little help in salvaging whatever reliable information may have been gathered in the course of the interview. Furthermore, when an inconsistency score is marginal, the user of the results is faced with some difficult choices about what and how much to believe, and again, the overall score is of no real help. Also, overall consistency/validity scales are not designed to detect inappropriate answers to isolated but perhaps crucial items. A low inconsistency score does not mean that all responses are valid.

Redundancy is clearly a valuable measure for dealing with errors that are approximately random across items, but it too has its drawbacks. The appropriate amount of redundancy that a test form should have is to some extent subject specific; a given amount of redundancy in a fixed-length questionnaire may lead to boredom or other negative reactions in some interviewees, yet the same amount of redundancy may be insufficient when the same form is administered to, say, a person who has poor language skills or cognitive impairment.

In the past, we have accepted the limitations and drawbacks in our data gathering techniques because our alternatives were severely limited. With computer technology, however, we may be able to create some new alternatives. There are two general approaches to dealing with data validity problems that are possible with computer technology, but not with paper-and-pencil technology.

THE PASSIVE APPROACH: ANCILLARY DATA

One obvious difference between a computer and a piece of paper is that the computer is capable of recording a wide range of ancillary information along with the interviewee's responses. Timing of response latencies, measurement of motion by ultrasonic detectors, and measurement of the force with which a response key is pressed can all be accomplished unobtrusively using existing hardware. Observations of eye movements and recording of data from skin electrodes are also possible, but any system incorporating overt physiological monitoring is likely to be perceived by the interviewees as a "lie detector," and hence, the use of such systems will probably be limited to special situations. The crucial advantage of these kinds of ancillary data is that they can be recorded for every response, thus making it possible in principle to detect relatively isolated problems as they arise during the course of a data gathering session. It is not proposed that ancillary data will provide an unambiguous indication every time there is a problem during an interview or that these measures will supplant traditional consistency and other checks. Rather, the ancillary information in combination with more tradi-

tional consistency/validity checks should enhance significantly the likelihood of being able to localize invalid responses. In some instances, such as when an interviewee's response latency abruptly drops from 3 ± 1.5 sec to .5 ± .2 sec, the ancillary data alone would be sufficient to diagnose the problem and identify the problematic responses, but, of course, these easy cases represent only a fraction of the total. Nonetheless, the potential gain from having item-by-item ancillary data is profound; with such data, it is feasible at least to consider separating valid from invalid data when a moderate level of inconsistency is encountered.

Ancillary data may be of interest for reasons other than error detection. A change in behavior from one item to another could arise as a result of one item's having, say, unusual emotional significance for the interviewee. A recently widowed interviewee might react strongly to questions about losses and loneliness; an interviewee who cannot stand his/her spouse might also react strongly to the same items, but for rather different reasons. In some settings, the affect associated with a given question may be of considerable interest, but in other settings, emotional responses may be regarded merely as a source of noise in the ancillary data.

Much research is needed to establish that the kinds of ancillary data that have been discussed are of any value at all in detecting substantively interesting events, whether these events are errors or emotional responses. A pilot study was conducted to explore the feasibility of using response latency data to detect problems or other significant events during a computer interview. The results, although limited, illustrate some of the problems of, as well as the potential gains from, gathering ancillary data during interactive interviewing.

Method

Eleven patients from the inpatient and partial hospital services of Butler Hospital were referred by their physicians for computer interviewing. The subjects, nine females and two males, ranged in age from 24 to 65 years, with a median of 53 years. The diagnoses (DSM-III) included affective psychoses (four cases), schizophrenia (three, including two paranoid schizophrenics), temporal lobe epilepsy, neurotic depression, personality disorder, and depressive reaction with personality disorder (one each). The subjects were not selected to be representative of any specific population. A computer interview developed for other research purposes (L. Hay, Note 1) was administered to the subjects. The interview was designed to produce a comprehensive patient problem inventory, and it covered role performance, role relationships, cognitive symptomatology, beliefs, affect, physical problems, environmental problems, and other problematic behaviors. All items were true/false. The items were presented on a video terminal attached to a PDP-10 timesharing system, which recorded responses and the latency between question presentation and the interviewee's first pressing of a key in response.

Following each interview, a questionnaire was given to each subject to ascertain the interviewee's subjective reactions to the computer interview. In order to ascertain whether unusually long or short response latencies might be associated with validity-threatening or other special events, a second questionnaire was given to each subject 1 day after the computer interview. This computer-generated questionnaire contained 12 items the subject answered on the computer interview, the 6 items having the shortest response latencies and the 6 items having the longest. The 12 items were, of course, different for each subject. For each of the 12 items, each subject rated his/her impression of the intelligibility of the item, the affective salience of that question for him/her personally, and whether or not he/she felt any reluctance to answer the question.

Results

The duration of the computer interviews ranged from 40 to 91 min, with a median of 62 min. Because of branching, not all subjects answered the same number of items. Role eligibility items and certain follow-up probe questions were excluded from the analysis because they were small in number and were qualitatively different from the screening questions that constituted the bulk of the interview. Also, a background section at the beginning of the interview was omitted for similar reasons. The number of responses analyzed for each subject ranged from 164 to 213, with a median of 188. Median latencies varied across subjects from 1.30 to 7.35 sec, with an overall median of 3.08 sec. The latency distributions were all positively skewed, and most subjects had a small number of extremely long latencies (up to 5 min). After preliminary analyses, latencies in excess of 60 sec were discarded from the data; this procedure had a negligible impact on the major findings.

As in most other studies, the interviewees reported the experience of being interviewed by a computer to be pleasant; 6 of the 11 rated the experience as "very pleasant," and none found it unpleasant.

The results from the questionnaires comparing long- vs. short-latency items were negative; there were no trends suggesting that the most extreme long-latency items were harder to understand, more affectively significant, or more personally intrusive or threatening than the extreme short-latency items. In part, the negative results from the follow-up questionnaire can be attributed to the tendency for subjects to show strong response biases on the follow-up questionnaire (e.g., answering that all 12 questions ask about an issue about which he/she has strong feelings, or asserting that all the questions were easy to understand); but even when there was within-subject variability, there was no evidence that response latency was strongly related to item difficulty, the affective significance of the item, or the personal intrusiveness of the item.

Evidently, the most extreme latencies from a long interview are primarily the result of factors not measured in the follow-up questionnaire or the methodology used was not adequate to detect the postulated effects. At times, interviewees did stop to ask the research assistant who was present in the room to explain a question or to make a comment about the interview. In future studies, special efforts should be made to record the nature and time of such events. More

Table 1
Latency Rankings of Computer Interview Sections Within Subjects

Questionnaire Section*	Subject 1	2	3	4	5	6	7	8	9	10	11
Work Role Performance	20	4	19	19	14	21	21	21	15	14	21
Leisure Activities	19	16	21	21	10	20	20	20	19	17	20
Household Role Performance	18	20	16	20	20	13	10	5	20	21	11
Money Management	7	21	20	9	17	15	18	12	12	19	17
Environmental Problems	2	9	11	10	11	19	14	3	7	9	12
Prior Treatment and Compliance	10	12	13	15	15	12	2	16	14	20	16
Major Social Relationships	5	10	15	18	16	14	7	17	21	12	19
Desired Relationships	15	18	18	14	12	4	19	7	13	18	7
Sexual Problems	13	14	10	13	13	17	13	4	6	7	8
Social Behaviors	6	13	9	16	19	11	17	13	18	11	13
Mood and Affect	9	11	8	11	8	9	11	18	9	15	14
Suicide/Self-Harm	21	2	5	4	3	18	16	1	1	8	15
Anger and Aggression	12	17	17	1	18	5	8	2	8	16	6
Blunted/Inappropriate Affect	17	19	14	12	21	16	15	9	16	13	18
Sleep and Appetite	3	5	6	2	1	3	4	14	2	6	4
Alcohol and Drug Abuse	4	1	12	3	9	7	9	8	10	2	2
Compulsive Behaviors	14	8	1	17	7	10	1	15	17	4	1
Beliefs and Attitudes	16	7	2	5	6	6	5	11	11	3	5
Memory and Cognition	11	15	7	6	4	2	12	10	5	10	10
Illusions/Hallucinations	8	3	3	8	2	1	6	6	3	1	3
Physical Disorders	1	6	4	7	5	8	3	19	4	5	9

Note–Rank 1 = shortest latency; Rank 21 = longest. **In order of appearance.*

powerful studies with human observation of interviewee behavior and better follow-up probe techniques are needed to study the causes of variation in item-by-item latencies. Even if, however, the latencies for individual items are not reliable indicators of a substantively interesting event, there remains the possibility that a consistent increase or decrease in latency across a group of related items might have substantive implications. This consideration led to an analysis of the variation in latency across blocks of related items within each subject. The items in the interview were divided into 21 a priori groups on the basis of content. Within each subject, latencies were ranked and a Kruskal-Wallis one-way nonparametric analysis of variance was done comparing the 21 item groups. For all subjects, it was found that there were statistically significant differences in latency ranks across item groups; significance levels for the Kruskal-Wallis test with 20 degrees of freedom ranged from .0121 to .0001. Table 1 gives the rank order of the 21 sections for each subject, as determined by the mean latency rank.

The mere existence of significant differences in latency across item groups does not, of course, imply that these latency differences have any substantive significance. The techniques of exploratory data analysis (Tukey, 1977) were used to examine the latency data for clues as to the substantive significance of the variations. A nonlinear data smoother was used to estimate a first approximation of the time trend in latency over the course of each interview. Nonlinear smoothing is a robust technique based on running medians that is relatively unaffected by isolated extreme values in a data sequence but will respond to trends that are consistent across adjacent data points. The particular smoothing algorithm used was "4253H, twice" (Velleman, Note 2). Scatterplots of item latency as a function of question sequence for Subjects 2 and 5 are shown in Figures 1 and 2. The solid curve is the time trend as estimated by the nonlinear smoother. Latencies greater than 21 sec are plotted as small circles along the top of the figures.

One feature displayed by the time trends for almost all subjects is a serial position effect; within each subject, latencies are generally higher at the beginning of the interview than at the end. In Table 1, it can be observed that the sections at the beginning of the interview tend to have high rankings, whereas those at the end tend to have low rankings. In one subject, Subject 8, there may also be a general upturn toward the end of the interview. In addition to the serial position effect, there

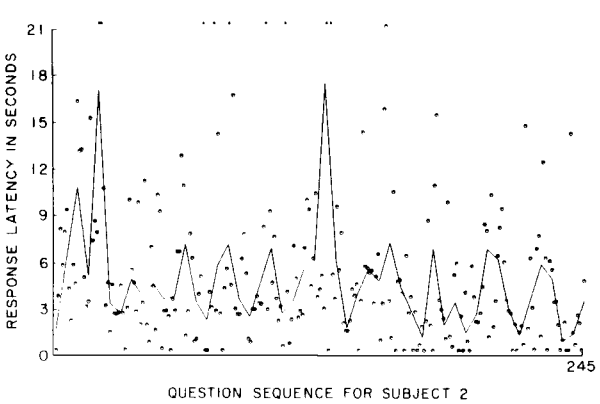

Figure 1.

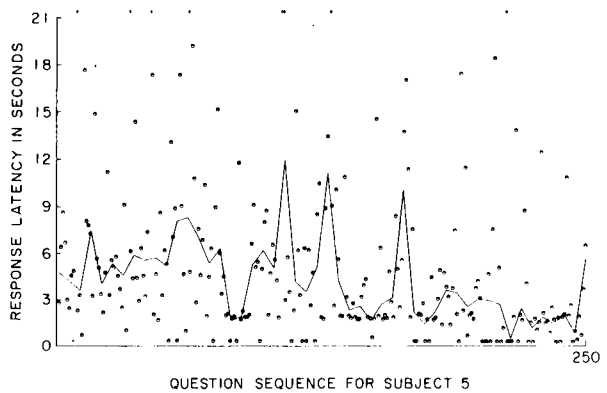

Figure 2.

is evidence that some portions of the questionnaire are associated with relatively long latencies across subjects, possibly because of item difficulty. The section on blunted or inappropriate affect, for example, seems generally to be associated with longer latencies than are the surrounding sections.

Beyond these general effects, however, there remains considerable variation across subjects with regard to which sections have high or low rankings. Certain portions of the questionnaire, particularly those on suicide, anger, and compulsiveness, seem to have especially large subject-to-subject variability. The substantive significance, if any, of these apparent intersubject differences must, however, be established empirically. When one observes in Table 1 that, for example, Subject 1 took a seemingly unusual amount of time to respond to the questions on suicide and self-harm, it is tempting to make an inference about the clinical significance of the questions in that section for that individual and/or the validity of the responses. Other aspects of the data suggest, however, that the relationship, if any, between response times and clinical importance is far from simple.

Table 2
Relation Between Reported Presence or Absence of Problem Indicator and Response Latency

	Problem Indicator					
	Absent		Present			
Subject	I	L	I	L	H	p
1	45	1.18	130	1.72	5.43	.0189
2	185	2.77	28	5.48	10.71	.0012
3	114	2.73	82	3.46	1.35	.2434
4	89	6.95	120	7.61	.42	.5219
5	104	3.96	82	3.01	.08	.7744
6	94	3.92	83	2.73	8.16	.0044
7	153	1.43	50	1.95	4.87	.0258
8	178	2.72	31	5.82	18.84	.0001
9	120	3.37	68	3.52	1.01	.3166
10	122	2.85	42	6.07	8.34	.0040
11	131	2.97	53	3.20	.99	.3205

Note–I = number of items. L = median latency (in seconds). H is the value of the Kruskal-Wallis statistic (df = 1). The p values are two-tailed.

Within each subject, a Kruskal-Wallis test was done, comparing the response latencies for items on which the subject's response indicated that a potential problem was present with items for which the subject indicated that the symptom or behavior was not present. The results are summarized in Table 2. For 5 of the 11 subjects, latencies were found to be significantly longer by a two-tailed test when the response was that the problem indicator was present. For five other subjects, there was no significant difference in latency between problem indicator present vs. absent items, and for one subject, there was a significant trend for the problem-absent items to be associated with longer latencies. The existence of individual differences of this magnitude complicates the problem of constructing reliable methods for the detection of significant events during an interview.

Discussion

It would be inappropriate to draw too many conclusions from this study of a small and unrepresentative sample. One conclusion that can reasonably be drawn, however, is that latencies are complexly determined and that considerable research will be needed to ascertain what "signatures" in latency patterns are associated with validity-threatening or other events in different categories of individuals. Nonetheless, even though large individual differences and major sources of noise variance pose difficult challenges, the approach of recording ancillary data during an interview still promises some important advantages. At a minimum, we can devise systems to detect instances of gross misbehavior on the part of the interviewee, such as the interviewee's beginning to respond randomly at the highest possible speed to finish the interview quickly. With somewhat more sophistication, it should be possible to detect the onset of interviewee fatigue when a long instrument is being administered. Even if it is not possible to determine with a great degree of confidence that the response to any single item is or is not valid, it should still be possible in some instances to determine that the responses to a given group of items or responses in a given content area seem to be relatively less reliable than other responses. Such a capability would represent a major improvement over whole-questionnaire validity scales. It should be noted that a computer system that records ancillary data along with the responses to a standard form always provides information that is at least as good as that which is obtained from the same instrument administered as a paper-and-pencil form. The computer always produces the same scale values that are computed from a paper form, and the ancillary information will at least sometimes provide further information of value.

Nonetheless, the ancillary-data approach does not represent the best that can possibly be done. Although the examination of ancillary data after the completion

of an interview may provide evidence about the nature and extent of validity problems, it still can help only in discarding unreliable information; it does not directly aid in obtaining valid answers. With this approach, the computer is still left in the same role as a sheet of paper: a passive medium for recording information.

ACTIVE APPROACHES

The computer has the power not simply to gather information but also to act upon it. A computer system that can detect possible inconsistencies or other problems should be able to take measures to attempt to correct the situation. To see what a computer system with an active approach to dealing with data gathering problems might do, we must turn our attention to the third data gathering technology currently in clinical use, namely, the human interviewer.

A human clinical interviewer going through a structured interview with a client is processing information on at least two levels simultaneously. On one level, the interviewer is going through an interview schedule with the goal of obtaining answers to the questions in that schedule. On this level, the human interviewer, the computer, and a paper form differ primarily in the medium by which the questions are presented to the interviewee. At the same time, however, the human interviewer is also making ancillary observations about response latencies, movements, facial expressions, tone of voice, and other variables. The spectrum of ancillary data that a human can observe vastly exceeds that which a computer system might monitor, but, of course, the human is not as systematic and thorough in his/her data collection as a computer would be. The human interviewer processes the responses and ancillary information for evidence that a given response may be invalid or unusually significant. If the circumstances warrant doing so, the interviewer can interrupt the lower level interviewing task (i.e., the task of going through the schedule) and stop to explain a question more thoroughly, ask the question in a somewhat different way, ask if the interviewee is fatigued or distressed, enjoin or cajole the interviewee to cooperate, or intervene in any number of other ways. The observational powers and interventional flexibility of a human interviewer make this data gathering technology the most powerful available, in the sense that a human interviewer can obtain relatively reliable information from many interviewees who could not or would not provide reliable responses on a paper form. The human interviewer technology, like the other two, has drawbacks also. Skilled human interviewers are expensive to train and use, and their behavior is not always as standardized as one might prefer.

In order to give the computer some of the observational power and interventional capabilities of the human interviewer, it will be necessary to redesign radically our current computer interviewing systems. As mentioned above, a human interviewer processes information on two levels simultaneously, the lower level being the nominal interview and the higher level being a metainterview. That is, the higher level is concerned with questions about the lower level interview, such as: "Did the interviewee understand the last question?"; "Is the interviewee being evasive?"; and "Is this person too upset to continue?" In order to carry out such a multilevel information processing task on a computer, it will be necessary to replace the current interviewing programs with a two-level executive program. The lower level of this executive program would be a questionnaire driver like those now in use, and the higher level component would be responsible for analyzing the interviewee's responses and ancillary data for evidence of unusual events and for selecting a suitable intervention should it appear likely that some event that might affect the interpretation of the responses has occurred. This higher level component might be called an interviewee behavior assessment monitor (IBAM). Even though an IBAM would not be able to employ many of the interventions that are possible for a human interviewer, there are still many steps that it could take. Among many other possibilities, and IBAM could: (1) insert extra items into the normal interview to check on the reliability of earlier responses, (2) insert metaquestions about whether the interviewee has particularly strong feelings about the topic of an earlier question, (3) recommend that the interviewee stop to rest, (4) enjoin the interviewee to cooperate more fully and back up the interview to the point at which the data first seemed to become unreliable, (5) call a human interviewer in from another room to handle problems that seem beyond its capabilities, and (6) switch to an alternate form of the interview if the interviewee seems to be having difficulty completing the standard form.

A recently described computer interviewing system incorporates a limited set of interventional features similar to some of those described above (Johnson, Giannetti, & Williams, 1979).

A variety of IBAMs might be necessary to handle different kinds of interviewees; different intervention strategies are appropriate for cognitively impaired interviewees and for uncooperative or emotionally distressed interviewees. The content and purpose of the basic questionnaire or test would also affect intervention strategies.

An active, interventionist approach such as the IBAM seems to be the only satisfactory way of overcoming the fundamental drawback of passive data gathering systems, namely, that however much ancillary data one gathers, there will be times when it will not be possible to determine after the interview what was happening when some crucial response was given. The best time to investigate a potential problem is right after it happens, and only a

system that can react to problem signals during an interview can initiate such an investigation and, if necessary, take corrective action. Thus, in a suitably designed active system, a low signal-to-noise ratio in the ancillary data is not as serious a problem as it is for a passive system.

CONCLUSION

Computer information gathering systems will never wholly replace human interviewers in clinical applications and may never wholly replace paper forms. We have the opportunity, however, to create a new technology with unique properties: one that will share the strengths both of standardized testing and of human interviewing. Before the first effective systems can be built, much research is needed on human behavior in computer interviewing situations, and formidable design and programming problems must be solved. Nevertheless, the potential benefits of these new approaches for clinical and research applications are such that we should not hesitate to face the challenges.

REFERENCE NOTES

1. Hay, L. The computer interview. *Post Newsletter,* 1(1). Providence, R.I: Butler Hospital, Department of Evaluation, February 1978.

2. Velleman, P. H. *Robust non-linear data smoothers: Definitions and recommendations* (Economic and Social Statistics Tech. Rep. 776/001). New York: Department of Economic and Social Statistics, School of Industrial and Labor Relations, Cornell University, 1976.

REFERENCES

ANGLE, H. V., ELLINWOOD, E. H., HAY, W. M., JOHNSEN, T., & HAY, L. R. Computer-aided interviewing in comprehensive behavioral assessment. *Behavior Therapy,* 1977, 8, 747-754.

BRADBURN, N. M., & SUDMAN, S. *Improving interview method and questionnaire design.* Washington, D.C: Jossey-Bass, 1979.

COLBY, M. Computer psychotherapists. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems.* Norwood, N.J: Ablex, 1980.

GREIST, J. H., & KLEIN, M. H. Computer programs for patients, clinicians, and researchers in psychiatry. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems.* Norwood, N.J: Ablex, 1980.

HAY, W. M., HAY, L. R., ANGLE, H. V., & ELLINWOOD, E. H. Computerized behavioral assessment and the problem-oriented record. *International Journal of Mental Health,* 1977, 6(2), 49-63.

JOHNSON, J. H., GIANNETTI, R. A., & WILLIAMS, T. A. Psychological systems questionnaire: An objective personality test designed for on-line computer presentation, scoring, and interpretation. *Behavior Research Methods & Instrumentation,* 1979, 11, 257-260.

JOHNSON, J. H., & WILLIAMS, T. A. Using a microcomputer for on-line psychiatric assessment. *Behavior Research Methods & Instrumentation,* 1978, 10, 576-578.

JOHNSON, J. H., WILLIAMS, T. A., KLINGLER, D. E., & GIANNETTI, R. A. Interventional relevance and retrofit programming: Concepts for the improvement of clinician acceptance of computer-generated assessment reports. *Behavior Research Methods & Instrumentation,* 1977, 9, 123-132.

TUKEY, J. W. *Exploratory data analysis.* Reading, Mass Addison-Wesley, 1977.