SESSION VII COMPUTER-BASED TESTING AND MEASUREMENT

JOSEPH L. ZINNES, *Indiana University, President*

The effects of computerized experimentation on response variance

VICTOR REZMOVIC

Northwestern University, Evanston, Illinois 60201

The advantages of using computerized procedures to replace human experimenters are examined. An experiment comparing two methods of presenting personality inventories indicated that participating in a computerized experiment increases response variance in a subsequent testing situation. Some possible negative consequences of using the computer to replace the human researcher are also discussed.

Recent findings in social psychology have indicated that the behavior of subjects in laboratory experiments may be more a function of certain artifacts present in the experimental situation than a function of the independent variables (Rosenthal & Rosnow, 1969). The caveats proposed by researchers in this area have led to a reexamination of the methods used in the laboratory experiment. Among the more common problems discussed are biasing effects associated with the presence of a human experimenter.

Role of the Experimenter

The unintended influence of the experimenter has been pointed out by McGuigan (1963) and elaborated by Rosenthal (1966). Rosenthal's analysis examines two general categories of experimenter influence. The first, "experimenter effects," deals with biological, personal, or social attributes of the investigator that influence the subjects to respond as a function of those attributes. The second influence is through mechanisms associated with "experimenter expectancy effects." Experimenter expectancy involves the transmission of cues from the experimenter to the subject as a function of hypotheses that the experimenter may have about his research. Experimenter effects and experimenter

expectancy are necessary but not sufficient to produce biased responding. In order for these effects to work, the subject population must be susceptible, and the task and setting must be conducive to produce bias. The presence of the human experimenter may also elicit "evaluation apprehension" in the typical laboratory subject. The presence of evaluation apprehension involves the perceived power held by the experimenter which impels the subject to behave in a way that will result in a favorable evaluation.

Unintended performance variation by subjects may influence results in one of two ways. First, subject characteristics or environmental variables may act in a random manner. When this occurs, the meaningfulness of results is not threatened. This situation results in conservative statistical tests, since random error only increases the denominator of an F ratio, leaving the numerator, or difference between treatments, unaffected. Establishing a significant relationship probably means that, once the experiment is replicated with better control over extraneous variables, the relationship should be stronger.

The second influence occurs when performance variation by subjects is a systematic function of undesired variables. For example, in cases where treatments interact with subject variables, as in the case of Rosenberg's (1965) explanation of dissonance findings in terms of evaluation apprehension, or extraneous demand characteristics, interpretability of results is greatly threatened. The results of such interactions are the emergence of "significant" findings when none exist (false positive findings), or the failure to find significance (when present) because an

Much of this paper is based on research conducted as part of the author's master's thesis research, under the direction of Edward J. O'Connell at Syracuse University. Additional time on the manuscript was supported by Contract C-74-0115 from the National Institute of Education. Thanks to Evi Rezmovic for her helpful suggestions on preparation of the manuscript.

extraneous variable has decreased statistical power (false negative findings).

The Computer as Experimenter

One proposed solution to the problems outlined above has been to replace the human experimenter, wherever possible, with automated procedures that minimize contact with the human researcher. The advantages of using computers in experiments have been discussed by Cooperband (1966), Johnson (1967), Kleinmuntz and McLean (1968), Messick and Rapaport (1964), Smith (1963), and Stang and O'Connell (1974).

A major advantage in using the computer is that it brings a high degree of standardization to the experiment. The computer, unlike the human experimenter, is not influenced by changing moods or by the physical or psychological characteristics of the subjects being tested. By providing each experimental subject with a highly similar set of stimuli to which he/she must react, random variation due to irrelevant influences should be reduced.

The speed and memory of the computer should help the experimenter to generate more experimental designs in which stimulus presentation is contingent on the subject's previous response(s). Using the computer to assign subjects to treatments can protect the human researcher from unintended bias.

The computer provides the researcher with an excellent research assistant: It never gets bored with repetition, and it provides time for the researcher to think and create. With adequate facilities and preplanning, data may be analyzed as they are collected and the time needed to complete a research project can be greatly reduced.

One of the most important consequences of using the computer is that it may increase "experimental realism" (Aronson & Carlsmith, 1969). Presumably, subjects using a computer become more involved in the situation and improve their attitudes toward the experiment.

Although the proponents of computerized experimentation have elaborated many advantages to replacing the human experimenter with a mechanized counterpart, few discussions have considered possible negative side effects. It is not known, for example, whether new types of demand characteristics may be elicited from subject-computer interactions, or what disadvantages may be associated with reducing contact between a subject and a human researcher. Although subjects seem to enjoy computerized tasks, it may be, as Lee (1970) has suggested, that computers give the impression of being all-knowing and make some subjects anxious.

Personality researchers are aware that subjects strive to present themselves in such a way as to be evaluated favorably by the experimenter. Computerized methods of testing seem to hold the promise of reducing some of the "evaluation apprehension."

To date, several studies have compared computerized methods of administering personality instruments to the classical methods employing a human experimenter. Researchers in these studies have hypothesized that subjects would respond more openly in a computerized testing situation. Evan and Miller (1969) administered a 120-item questionnaire to two groups of subjects. The first group was presented with questions via computer Teletype. The second group received paper-and-pencil administration with a human experimenter presenting the materials and providing directions. The questionnaire contained items from the Allport-Vernon-Lindzey Scale of Individual Values (AVL), the K scale of the MMPI, items from the MMPI relating to manifest anxiety, the Srole Scale relating to perceived sociocultural anomie, and neutral items, that is, questions that elicited impersonal opinions and factual information. Evan and Miller's prediction that subjects in the automated situation would be more open (i.e., more willing to check as true items of a highly personal or embarrassing nature) was statistically significant for only the religious value items of the AVL ($p < .005$). Differences on the combined-scales score were not significant, due to the unanticipated results of group differences on the impersonal opinion items. With the neutral items excluded, the overall group differences were significant ($p < .034$). Although the study did find some significant differences between the computer and paper-and-pencil groups, the overall result did not provide strong support for their hypotheses. This may have been due to two weaknesses: (1) The subjects were MIT undergraduates, who, according to Evan and Miller, are a very "restricted, homogeneous, and atypical group"; and (2) although the authors attempted to control for prior experience with computers, their postexperimental data indicated that the groups did differ in computer-usage histories.

In a similar study, Koson, Kitchen, Kochen, and Stodolosky (1970) investigated subjects' responses to questions presented via: (1) a cathode ray tube (CRT), (2) a paper-and-pencil test, and (3) a human experimenter who read the questions to the subjects. Koson et al. (1970) used highly threatening MMPI items, the K scale of the MMPI, and neutral items. Although their results were in the predicted direction, they were not significant. The authors concluded that "perhaps there wasn't enough 'computerness' in the stimulus value of the cathode ray tube and push buttons." This criticism may also be made of Evan and Miller, who used the computer mainly as an elaborate presentation device. By not using the full interactive capabilities of the computer, the resulting situations may have been low in "experimental realism."

Both studies used a number of items of a highly threatening nature. However, the only subset of items that produced significant differences between groups was the AVL religious value items. The fact that no

differences were found on the threatening items may suggest that such stimuli are affected by evaluation apprehension in most contexts. Items that contain a moderate amount of threat, such as items relating to religious beliefs, may be more influenced by environmental cues, such as presence or absence of the experimenter.

METHOD

The present study utilized two frequently used personality inventories: Rotter's (1966) Internal-External Locus of Control Scale, and Crowne and Marlowe's (1960) Social Desirability Scale. Both scales have been extensively used in past research, and score norms as well as validity and reliability data are available (Robinson & Shaver, 1973). Neither scale contains items that are considered highly threatening or extremely personal.

The two scales were administered under two different methods at two different times. In the first method (PP), a human experimenter presented the materials as a paper-and-pencil task.¹ The second method (COMP) used the computer to present the stimulus materials. Half of the subjects (Group 1) responded to the two scales in the presence of a human experimenter and then returned several weeks later to complete the same scales administered by the computer. The other half of the subjects (Group 2) received the computerized testing first and were retested in the presence of a human experimenter when they returned a second time.

It was hypothesized that computerized testing would produce (1) differences between group means, or (2) increased variance in responses. Subjects at the extremes were expected to become more extreme as a consequence of interacting with the computer.

Subjects

The subjects were 70 male and 28 female students in the introductory psychology course at Syracuse University. Subjects were chosen at random from a large subject pool and were randomly assigned to Group 1 or Group 2.

Computer Facilities

The computer used to present the stimulus materials was a DEC PDP-8/F with 8K core, with a Model 33 Teletype.² An interactive program was written in FOCAL-8 to present the SD and I-E items and to record the subjects' responses.

RESULTS

Results of the study are presented in Table 1. Univariate analysis of variance showed no significant differences in the means between the groups. Com-

Table 1
Means and Variances of SD and I-E Scores in the PP Condition and the COMP Condition for Groups 1 and 2

Group	PP Condition		COMP Condition	
	SD	I-E	SD	I-E
1	13.4490*	11.3878	13.2245	11.0204
	21.1453**	18.5227	30.2148	24.0198
2	14.2653	10.6531	14.2857	11.0408
	37.9505	29.6556	28.5712	22.5293

*Means

**Variances

parisons of the variances, however, indicated that there was a significant difference in variance on both the SD and I-E in the PP condition, with Group 2 variance significantly higher.³ (For Group 1, the PP condition was the first experimental session; for Group 2, the PP condition followed participation in the COMP condition.) Increased variance was attributed to interaction with the computer.

To insure that the significant difference in variance was not due to more experience with the testing situation on the part of Group 2, the same tests were conducted comparing the two groups on the COMP condition alone. No significant differences were found.

DISCUSSION

Method of presentation did not significantly influence the mean scores of the two groups in this experiment, nor did it appear to influence all members of a group equally. The results indicated that interacting with the computer increased the variance of scores on subsequent testing, indicating that some subjects did respond in a more extreme manner at both ends of the scale. The experience of using the computer did not influence variance directly, but served as a "releaser" for subjects' performance on the second visit to the laboratory.

Introducing a computerized assessment procedure into the laboratory may have many beneficial effects, but the presence of the computer may also interact with many other variables, for example, the nature of the stimulus materials. In the area of personality measurement, items that are highly threatening may be answered with equal uniformity; those items perceived as less threatening may bring about greater variation in responses.

Recent research by Johnson and Baker (1973) in the area of ability measurement indicated that the introduction of a computerized testing procedure increased response variance. Ability level, however, interacted with method of testing, and the increased variance was due to low-ability subjects, who did not perform as well when tested by the computer. Implications of the study by Johnson and Baker are that computerized testing may have detrimental effects on certain subjects, and that automated procedures may not always be desirable. The presence of a human experimenter may be reassuring to some experimental subjects in conditions that involve the measurement of their abilities.

If the full interactive capabilities of the computer can be put to use, the subject's involvement should be high, and willingness to reveal socially undesirable material may increase. One way to achieve this is to program the interaction between computer and subject so that it resembles the dialogue between the subject and a human experimenter. Individual differences

may also be important. The attitude of a subject toward automation, and/or previous experience with computers and knowledge of computer functions may also determine the degree to which a subject feels comfortable in his interaction with the computer.

In general, the laboratory situation is a complex interaction between the subject's characteristics and the environmental features. As the computer becomes a more standard feature in the research process, its advantages and limitations should be more clearly specified.

REFERENCES

- ARONSON, E., & CARLSMITH, J. M. Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. II). Reading, Mass: 1969.
- COOPERBAND, A. S. The use of a computer in conducting psychological experiments. *Behavioral Science*, 1966, 11, 307-311.
- CROWNE, D., & MARLOWE, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, 24, 349-354.
- EVAN, W. M., & MILLER, J. R. Differential effects on response bias of computer vs conventional administration of a social science questionnaire: An exploratory methodological experiment. *Behavioral Science*, 1969, 14, 216-227.
- HAYS, W. L. *Statistics for the social sciences*. New York: Holt, Rinehart, & Winston, 1973.
- JOHNSON, E. S. The computer as experimenter. *Behavioral Science*, 1967, 12, 484-489.
- JOHNSON, E. S., & BAKER, R. F. The computer as experimenter: New results. *Behavioral Science*, 1973, 18, 377-385.
- KLEINMUNTZ, B., & McLEAN, R. S. Diagnostic interviewing by digital computer. *Behavioral Science*, 1968, 13, 75-80.
- KOSON, D., KITCHEN, C., KOCHEN, M., & STODOLOSKY, D. Psychological testing by computer: Effect on response bias. *Educational and Psychological Measurement*, 1970, 30, 803-810.
- LEE, R. S. Social attitudes and the computer revolution. *Public Opinion Quarterly*, 1970, 34, 53-59.
- MCGUIGAN, F. J. The experimenter: A neglected stimulus object. *Psychological Bulletin*, 1963, 60, 421-428.
- MESSICK, D. M., & RAPAPORT, A. Computer controlled experiments in psychology. *Behavioral Science*, 1964, 9, 378-382.
- ROBINSON, J. P., & SHAVER, P. R. *Measures of psychological attitudes*. Ann Arbor, Mich: Institute For Social Research, 1973.
- ROSENBERG, M. J. When dissonance fails: On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology*, 1965, 1, 28-42.
- ROSENTHAL, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.
- ROSENTHAL, R., & ROSNOW, R. L. *Artifact in behavioral research*. New York: Academic Press, 1969.
- ROTTER, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(Whole No. 609).
- SMITH, R. E. Examination by computer. *Behavioral Science*, 1963, 8, 76-79.
- STANG, J., & O'CONNELL, E. J. The computer as experimenter in social psychological research. *Behavior Research Methods & Instrumentation*, 1974, 6, 223-231.

NOTES

1. The author served as the experimenter in this study.
2. The computer facilities were provided by National Science Foundation Grant GY 10100, awarded to Warren Klare.
3. The statistical test used was that described by Hays (1973) for comparing two variances (p. 449). For the SD, $F = 1.79$, $df = 48/48$, $p < .05$; for the I-E, $F = 1.61$, $df = 48/48$, $p < .05$.