EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1970, 30, 803-810.

PSYCHOLOGICAL TESTING BY COMPUTER: EFFECT ON RESPONSE BIAS

D. KOSON, C. KITCHEN, M. KOCHEN, AND D. STODOLOSKY
Mental Health Research Institute
University of Michigan

IT is easy to use computers for administering multiple-choice questionnaires. This can be done by displaying questions on a TV-type screen or on a teletypewriter in front of the respondent; the respondent keys in his response while the computer waits. It is also possible, though not easy, to administer, by computer, questionnaires calling for the respondent to generate and type sentences. In neither case would the computer be used efficiently unless there were enough respondents (or other tasks) to keep the machine from being neither idle nor congested unreasonably long. Using machines in such a time-sharing fashion trades certain desiderata for the increased cost due to the computer's having to keep track of the traffic. What are these desiderata?

It is reasonable to suppose that one advantage to be gained by the use of computers rather than conventional means to administer psychological tests is an increase in honesty or a reduction in response bias of answers to certain sensitive or ambiguous questions. Response bias has been defined as the tendency to distort answers, consciously or unconsciously. The work of Rosenthal (1963) has persuasively demonstrated that the experimenter's or tester's expectations significantly affect the test outcomes. While such expectations might be attributed to the designer of a test or to the computer programmer, it is possible to minimize such effects by careful design.

Respondents will sometimes try to help confirm the experimenter's hypotheses (Orne, 1962), to respond when doubtful, to agree rather than disagree (Cronbach, 1946), or to place themselves in a

favorable light (Edwards, 1957). If a respondent is doubtful about the confidentiality of his answers, he tends to bear his responses so as to protect his interests (Dunnette and Heneman, 1956).

Many procedures have been recommended to combat or control response sets. Cronbach (1964) for one, asserted that response sets are reduced by any procedure that increases the structuration of the test situation. Jackson (1967) suggested that if scales were developed with half of the items true and half false (or agree-disagree), the massive cumulative effect of acquiescence in undirectionally keyed or grossly imbalanced scales would be avoided. An alternative is to estimate set and content variance from the same set of items (Messick, 1961; Bock, 1964); to develop separate scales for set and content from the same domain of items (Jackson, 1967); or to use forced- or multiple-choice between equally socially desirable items.

Proposed here is another way of further structuring the test situation so as to reduce the variability of measurements and effects of certain sets due to characteristics of the tester.

When the content of a question is regarded by the respondent as highly personal or disturbing, he may not be prone to respond honestly (Stricker, 1963). Smith (1963) has suggested that there may be some very personal and embarrassing material that a person may not want to express to another person, but might be able to express to an impartial machine, without fear of negative evaluation. Such biases as carelessness, acquiescence, social desirability, and embarrassment or defensiveness may operate in psychological tests and questionnaires. This may be because the test or research situation is perceived as either embarrassing and disturbing or not confidential. Then a situation which either conveys a sense of anonymity or confidentiality, or reduces the embarrassment or defensiveness with which sensitive material is divulged could reduce such biases. It is the matter of embarrassing or disturbing questions to which the present study addresses itself.

Major hypotheses are that subjects responding to computer-administered tests of threatening disturbing items will respond with less response bias and less defensiveness than subjects under paper and pencil administration, who in turn, will have less response bias and less defensiveness than subjects in an interview situation.

Further, emotionally neutral and impersonal questions will produce no significantly different response tendencies between the three conditions.

Method

Subjects

Sixty-eight University of Michigan undergraduates and graduate students volunteered as paid *S*s. These *S*s ranged in age from 18 to 26, the median age being 20. Males and females were used in equal numbers in all groups throughout the experiment. Twenty *S*s were randomly selected to rate potential questionnaire items independently of the experimental groups. The remaining 48 *S*s were divided randomly into three groups of 16 until each group contained eight males and eight females.

Materials

We attempted to construct a questionnaire containing three categories of questions. The first group of questions would be selected so as to be considered threatening or embarrassing to *S*s. The second group of questions would be selected to be emotionally neutral or nonthreatening. The third group of questions were from the *K*-scale items, taken from the MMPI. These items are "obvious" in the sense that the probability of being true is very high, and respondents perceive this. These items were designed to elicit denial and defensiveness.

Twenty *S*s were presented with the MMPI plus 10 additional questions supplied by the authors and asked to rate each question from one (neutral, nonthreatening) to five (threatening, embarrassing). Instructions to *S*s were in part, ". . . imagine responding to each question as if another person were asking it. Then ask yourself whether this item would embarrass you or inhibit you from answering truthfully. We ask that you grade your reaction to each item in terms of *how* threatening and *how* embarrassing it is to you."

The 11 questions receiving the highest mean ratings were used in our questionnaire as the "Threat scale." These ratings are closely akin to Social Desirability ratings (Edwards, 1957), and questions with high ratings, by virtue of their construction, are intended to elicit response bias. These items chosen as most "threat-

ening" or embarrassing had a mean score of 2.6, ranging from 2.4 to 2.8. The neutral items were chosen from a pool of 80 items each having a mean score of one.

Threat scale items were keyed in the socially undesirable direction. Response bias was operationally defined as the tendency to respond by means of the socially desirable answer rather than in response to the content of a question. For three randomly chosen groups this tendency and, operationally, the number of keyed answers should be the same. *K*-scale items likewise were keyed in the "obvious" direction. Neutral items were keyed randomly. The questionnaire, in its final form, contained 33 items, 11 each of threatening, neutral, and *K*-scale items, and were randomly mixed. Finally, with the help of some negations and reflections, half the items on each scale were keyed "true," and half "false," so as not to confound acquiescence with Social Desirability response bias.

Procedure

All three experimental groups received the same treatment in the first part of the experiment where they were asked to rate 10 items on the same 1–5 scale used by the original raters. These items included five items appearing on the questionnaire along with five other items chosen because of their demonstrated "threat" value. The rating session was conducted to see whether our *S*s in fact perceived as "embarrassing or sensitive," the questions used in our Threat scale. We expected the experimental effect to be more pronounced in *S*s if they rated items as more "threatening" than other *S*s. The second part of the experiment consisted of taking the questionnaire. One group of *S*s responded to the questionnaire by marking "T" or "F" to the appropriate question in the presence of *E*.[1] This is the paper and pencil group. The second group responded to the questions verbally after the question was read by the *E*. The *E* marked the questions instead of *S*. The third group responded to the questions as they were projected on the screen of a cathode-ray tube (CRT) by pressing the appropriate levers marked "True" or "False." The CRT was connected to a DEC PDP-8[2] which controlled the presentation of the questions. *E* was not

[1] We wish to thank Miss Valarie Bunce for serving as *E* in all the experimental groups.

[2] We are indebted to Mr. Peter Headly for his programming help.

present when the questions were answered. All three groups had a debriefing period with *E* in which they were asked the purpose of the experiment. In addition, the computer group was asked a question as to their experience with the computer. All *S*s were run individually and *E* was given instruction to standardize her behavior as much as possible. The mean number of answers per *S* in the keyed direction was tabulated for each group, by sex, on the Threat, Neutral, and *K*-scales. Total number of "true" responses was tabulated for each group, by sex. Mean scale score was computed for each group, by sex, for the threat rating scale.

Results

Table 1 presents the means and standard deviations for experimental and control groups based on mean number of keyed answers per *S* on the Threat-scale, and the means for each group by sex.

TABLE 1

Mean, "Honesty" Scores, Threat Scale

	Computer	Pencil-Paper	Interview
mean	3.13	2.56	2.31
s.d.	2.4	1.8	1.8
mean	*M* *F*	*M* *F*	*M* *F*
	2.50 3.75	2.63 2.50	2.50 2.13

Table 2 shows the same information for the *K*-scale.

The computer group answered more questions in the keyed direction and showed less defensiveness than control groups, based on higher Threat- and *K*-scale scores. An analysis of variance was

TABLE 2

Mean "Defensiveness" Scores, K-Scale

	Computer	Pencil-Paper	Interview
mean	6.70	6.56	6.00
s.d.	2.6	2.0	2.1
mean	*M* *F*	*M* *F*	*M* *F*
	5.88 7.50	6.75 6.00	6.25 5.75

applied to total Threat-scale for each group, and sex. The results yielded an *F* of .625, which indicate that the differences are not significant.

Table 3 shows total number of keyed Neutral answers for each group. There were no significant differences between groups.

TABLE 3

Neutral-scale Scores

	Computer	Pencil-Paper	Interview
Total	81	82	80

Table 4 shows the mean threat-rating score for each group by sex. There were no significant differences between groups but the computer group shows a large difference between sexes. Threat-rating score and threat-scale "honesty" score correlation was essentially zero, that between threat-rating score and *K*-scale was +.14. The correlation between Threat-scale and *K*-scale scores for all *S*s was +.6.

TABLE 4

Threat-rating Score

	Computer		Pencil-Paper		Interview	
	M	*F*	*M*	*F*	*M*	*F*
mean score Per *S*	1.4	2.0	1.8	1.9	1.7	1.7

Table 5 shows the mean proportion of "true" responses for each subject by group and sex. Computer-group females tended to answer "true" more than males, but there were no significant differences between groups.

TABLE 5

Proportion of "True" Responses

	Computer		Pencil-Paper		Interview	
	M	*F*	*M*	*F*	*M*	*F*
mean score Per *S*	20.4	22.5	20.4	20.2	20.7	18.4

Discussion

Although the computer group scored higher on the Threat- and *K*-scales than the pencil and paper group, who, in turn, scored

higher than the interview group, the differences were not significant. The Threat- and *K*-scales correlated +.60 but since this is not very high and the *K*-scale differences were not large, an analysis of covariance yielded equally inconclusive results and was not reported. The difference between the computer group and the others seems largely due to the females in the computer groups, who earned high scores on the Threat- and *K*-scales. The difference between males and females in the computer group was significant at the .1 level (*t*-test) which in a preliminary study is small enough to warrant further study.

As expected, there were no differences between any of the groups on the Neutral-scale, a consequence of the fact that the probability of a certain number of the neutral questions being true of an *S* was the same for all groups. The overall proportion of questions answered in the keyed direction was .46 and variances for each group were similar and low. The proportion of questions answered in the keyed direction on the *K*-scale, however, was .58, highest of all scales, and the variances for all groups was uniformly high. This indicated to us that the *K*-scale items were not as subtle as intended. That is, the keyed direction was too obvious. On the Threat-scale, however, the proportion of keyed items answered was lowest, .24, and variances fell between Neutral- and *K*-scale. Here too, the keyed direction was fairly apparent, but the items were sensitive.

Since the variances were so high in this study, another study must reduce some of the "noise" which enlarges error terms. A more refined study should employ a test of defensiveness either matched across groups or as a factor. More importantly though, in the computer group a variety of attitudes were operating to generate the variability which we found, particularly between the sexes. A further study should systematically relate attitudes to performance in man-computer interactions.

In regard to attitudes and perceptual set, perhaps there wasn't enough "computerness" in the stimulus value of the cathode ray tube and push buttons. Perhaps the stimulus value of the computer setting could be altered to give more "computerness" to the setting, thus increasing the range of our independent variable.

The concept of "response bias" must be explicated. It could be defined operationally by responses to sensitive questions whose

answers are already known, as in studies of medical and social interviewing.

REFERENCES

Bock, R. D. Components of variance due to content and acquiescence in the *My* and *Pt* scales of the MMPI. Chapel Hill, N. C.: Psychometric Laboratory Research Memo No. 21, 1964.

Cronbach, L. E. Response sets and test validity. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 6, 475–495.

Cronbach, L. E. Further evidence of response sets and test design. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1950, 10, 3–29.

Cronbach, L. E. *Essentials of Psychological Testing.* New York: Harper, 1960.

Dunnette, M. D. and Heneman, H. G., Jr. Influence of scale administration on employee attitude response. *Journal of Applied Psychology,* 1956, 40, 73–77.

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research.* New York: Dryden, 1957.

Heneman, H. G., Jr. and Yoder, D. Employee opinion survey by remote control. *Personnel Journal,* 1953, 32, 169–173.

Jackson, D. N. In Berg, I. A. (ed.) *Response Set in Personality Measurement.* Chicago: Aldine, 1967.

Messick, S. J. Separate set and content scores for personality and attitude scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 915–923.

Orne, M. On the social psychology of the psychological experiment. *American Psychologist,* 1962, 17, 776–783.

Pearlin, L. I. The appeals of anonymity in questionnaire responses. *Public Opinion Quarterly,* 1961, 25, 640–647.

Rosenthal, R. On the social psychology of the psychology experiment; the experimenter's hypothesis as unintended determinant of experimental results. *American Scientist,* 1963, 51, 268–283.

Smith, R. E. Examination by computer. *Behavioral Science,* 1963, 8, 76–79.

Stricker, L. J. Acquiescence and social desirability response styles, item characteristics, and conformity. *Psychological Reports,* 1963, 12, 319–341 (Monograph Suppl. 2–12).