Journal of Personality and Social Psychology
1966, Vol. 3, No. 1, 28–37

USE OF TWO DIFFERENT RESPONSE MODES AND REPEATED TESTINGS TO PREDICT SOCIAL CONFORMITY [1]

LEWIS R. GOLDBERG AND LEONARD G. RORER

University of Oregon and Oregon Research Institute *Oregon Research Institute, Eugene*

The usefulness of an economical method for measuring social conformity was explored. The effects upon yielding of (a) accurate as opposed to inaccurate information and (b) instructions to remember and reproduce previous responses as opposed to instructions to ignore previous responses were studied. In addition, this study explored the predictability of individual differences in yielding through the use of a personality inventory, examining the effects of repeated administrations of the inventory as well as the effects of having Ss predict the responses of others as opposed to having Ss personally respond to the items. Results suggested that this approach to the measurement of social conformity is fruitful, although individual differences in yielding were poorly predicted by the personality inventory under any of the experimental conditions.

This methodological study of the prediction of social conformity was designed so that a number of previously unrelated hypotheses could be tested simultaneously.

Measurement of Social Conformity

Many social critics have commented on the apparent tendency of individuals in contemporary society to conform to group opinion. One result of this social commentary has been the attempt to devise techniques whereby conforming behavior could be studied in the laboratory. Asch (1952), using an experimental procedure in which one subject and eight stooges made judgments concerning the comparative lengths of lines, was able to show that individuals faced with a unanimous consensus would often disregard the evidence of their own visual sensations in order to conform to the group judgment. Crutchfield (1955) mechanized Asch's procedure so that no stooges were required and a number of subjects could be tested at one session. In addition to perceptual judgments, Crutchfield included both statements of opinion and unsolvable problems among the stimuli to which he had his subjects respond. He was able to show that the more ambiguous the stimulus, the greater the tendency of subjects to conform to group opinion.

Tuddenham (e.g., 1958a, 1958b), in an integrated series of yielding and conformity studies, replicated much of the earlier work, and, in addition, compared the effects produced by the provision of accurate as opposed to inaccurate information. He found that whereas inaccurate information increased judgmental disparity, accurate information induced convergence. Fisher and Lubin (1958) investigated the effect of distance between the subject's position and the spurious report upon two measures of influence: movement, the amount which the subject changed from his original position, and conformity, the extent to which the subject yielded to the other's position. While these two indices produce identical results when the initial distance is the same for all subjects, Fisher and Lubin demonstrated that as distance increases, movement tends to increase while conformity tends to decrease.

While the techniques developed by Asch and Crutchfield and utilized by Tuddenham, Fisher and Lubin, and others are ingenious and produce dramatic results, they are also expensive and time consuming. Much more practical as a mass testing technique is one described more than 40 years ago by Moore

[1]The study was supported by National Science Foundation Grant G-25123. Data analyses were carried out through the facilities of the Western Data Processing Center at the University of California in Los Angeles and the Statistical Laboratory of the University of Oregon. The authors wish to thank Richard Hammersley, William Johnson, Jacob Kind, George McCarger, and Patricia Taylor, who worked as research assistants on this project.

(1921)—an extension of the original method of Münsterberg (1914). The procedure, utilized more recently by Hastorf and Piper (1951), requires only that a group of individuals respond to a questionnaire on more than one occasion. The first occasion serves to obtain data concerning the subject's unbiased answers to the questions. The second administration serves to estimate the extent to which the subject will change his responses when presented with either correct or incorrect information concerning the manner in which the rest of the group responded to the first administration of the questionnaire. Hastorf and Piper demonstrated that this technique produced substantial conformity to purported group opinion even when the subjects were instructed to try to remember their reponses to the first administration and give the exact same response on the second administration (i.e., even though respondents were instructed to remember and reproduce their responses from one administration of a questionnaire to another, they still were influenced by data pertaining to the responses of others).

With the aim of further establishing this technique of measuring social conformity, the present study attempted to replicate and extend the work of Hastorf and Piper (1951) and Tuddenham (1958a, 1958b) by systematically manipulating: (*a*) the presentation of accurate as opposed to inaccurate information and (*b*) instructions to remember previous responses as opposed to instructions to respond without regard for previous responses. In addition, the two different indices of conformity proposed by Fisher and Lubin (1958) were operationalized and compared.

Search for Nonfakable Inventory Items

Psychologists attempting to improve the validity of personality inventories have consistently been frustrated in their search for items that are sufficiently subtle that they are not fakable. Though the forced-choice inventory appeared to promise a solution to this problem, recent experimental evidence indicates that the forced-choice format neither solves the problem of faking (e.g., Dicken, 1959) nor increases predictive validity beyond that of more simple formats (e.g., Heilbrun, 1962; Scott, 1963). The problem of faking stems from the fact that some subjects may try to present a favorable impression of themselves while others try to respond more honestly. One possible solution to the problem is to change the basic task of the subject from that of reporting his own internal states to that of assessing as accurately as possible the beliefs or attitudes of others—a task supposedly not amenable to faking (e.g., Campbell, 1950). Instead of asking a respondent to evaluate an item as true or false of himself, one might ask him to indicate whether most people would answer true or false, or whether some selected group of subjects (e.g., successful executives) would endorse the item or not (Gordon, 1949). Messick (1960), on the basis of a factor-analytic study of social desirability ratings, suggested that the task of rating the social desirability of personality test items might be used to extract personological information (i.e., that knowledge of how an individual sees people in the world around him might provide important data relevant to the prediction of that individual's behavior).

Goldberg (1962) tested Messick's hypothesis in a relatively informal manner. Thirteen students in an undergraduate course in personality assessment carried out individual research projects in which they pitted the typical personal endorsement procedure against social desirability ratings as potential indicators of personality trait variance. Results indicated that in none of the 13 studies did social desirability ratings do a better job than the traditional endorsement procedure in differentiating criterion groups, though in 5 of the studies they fared about as well. Consequently, these 5 studies were later cross-validated. In all cases the shrinkage upon cross-validation was greater for the social desirability rating procedure than for the typical endorsement procedure.

Jackson (1961, 1964) used desirability ratings of 45 inventory items to predict social conformity. His instructions, similar to those used by Edwards (1957), asked the subjects to judge on a 9-point scale the desirability of a true response to a statement when the statement is applied to other people. Jackson found that the desirability judgments had a

slightly higher relationship with his criterion (.29) than did the personal endorsement responses (.22), and suggested that in faking situations the disparity in favor of desirability judgments might be even greater. However, when Loomis and Spilka (1963) replicated Jackson's study, they found that desirability judgments correlated −.15 with the criterion, although the validity of the endorsement responses (.26) was about the same as that found by Jackson. The present study was designed as a modification of those by Jackson and Loomis and Spilka, introducing a number of control conditions absent from these previous studies. The hypothesis tested was that behavioral criteria can be predicted from predictions of the responses of others as well as from the more traditional personal endorsement responses.

Effects of Repeated Testings

On the basis of research carried out by Fiske (1957a, 1957b) and Mitra and Fiske (1956), Howard (1964) has hypothesized that, as a personality inventory is administered to a group repeatedly, interpersonal variability increases, while intrapersonal variability decreases (i.e., that the responses made by an individual become more consistent and more differentiating). A series of related studies tends to substantiate the proposal (e.g., Howard & Diesenhaus, 1965). Since response instability and interindividual similarity may both act as constraints upon predictive validity (Goldberg, 1963), it follows that repeated testing should increase validity by simultaneously maximizing both intraindividual response stability and interindividual variability. The present study undertook a direct test of the hypothesis that later, as opposed to earlier, administrations of a personality inventory would yield higher indices of predictive validity.

Traditional psychometric theory would hold that any one administration of an inventory provides but one estimate of the respondent's true score on that inventory. Successively obtained scores will not be equal, but should be distributed around the respondent's true score, which, if it were known, would provide the highest validity coefficients. From this it follows that the more accurately the true score can be estimated, the more accurate should be the prediction. Since the variance of the mean of the scores from repeated administrations is less than that of the individual scores, the mean provides a more accurate estimate of the true score than does any one of the individual scores. Therefore, the validity coefficients based on the mean would be expected to be higher than those for any one administration. If the hypotheses of Howard are correct, however, it may be that earlier administrations are simply adding error variance, so that predictions based on later administrations would be more accurate than those based on the mean. The hypothesis that predictions based on the mean are more accurate than those based on any single administration was among those tested in the present study.

In summary, the present study sought to ascertain further the usefulness of an economical method of measuring social conformity. In order to do so, it explored the effects upon two indices of social conformity of accurate as opposed to inaccurate information and compared instructions to remember and reproduce previous responses with instructions to ignore previous responses. In addition, in an attempt to improve the prediction of social conformity from personality inventories, the study investigated the hypothesis that conformity can be predicted as accurately from subjects' predictions of others' responses to inventory items as from the traditional endorsements of the same items. A final aim of this study was to investigate the effects of repeated testings upon predictive validity, discovering whether later administrations of an inventory are more valid than earlier administrations on the one hand and the mean of all repeated administrations on the other.

METHOD

Subjects

From a freshman dormitory at the University of Oregon, 198 coeds volunteered to participate for pay in this study, and 157 completed all of the experimental procedures. The mean age of the subject group was 18.0 years, with a standard deviation of .4 year.

Instruments

Opinion Questionnaire. The criterion of social conformity was obtained by means of a double administration of the Opinion Questionnaire previously used by Hastorf and Piper (1951), Jackson (1964), and Loomis and Spilka (1963). The 45 questions included 15 dealing with economic policies (e.g., "During a period of shortage the government should subsidize public housing."), 15 covering educational policies (e.g., "Teachers should be hired and assigned to their positions by educational specialists rather than by local school boards."), and 15 covering more general personal values (e.g., "There are cases where pre-marital sexual relations are justified."). The subjects were asked to indicate the amount of their agreement or disagreement with each statement on a scale ranging from 1 (complete agreement), through 5 (uncertainty), to 9 (complete disagreement).

For the second administration of the Opinion Questionnaire, each of the 45 questions was followed by a number described as the average response given by the subjects on the first administration of the Questionnaire. For the 25 questions with the smallest dispersions of group ratings on the first administration, the reported mean value was obtained by shifting the actual mean 3 points toward whichever end of the scale was most distant. The mean values of the 20 questions with the largest dispersions were reported accurately. This procedure was essentially the same as that used by Jackson (1964) and Loomis and Spilka (1963). It is important to realize that the accurate versus spurious distinction is confounded with differences in original item dispersions, a confounding also common to the previous studies.

Two different sets of instructions were utilized for the second administration of the Opinion Questionnaire. Approximately one-half of the subjects were asked to try to remember the answers they had given to the Questionnaire on its first administration and to give exactly the same response on the second occasion. These *memory* instructions were virtually identical to those used by previous investigators. The other half of the subjects were asked to fill out the Questionnaire "as you feel today, as if you had never filled out the Questionnaire before" (*current* instructions). Both the memory and current groups filled out the Questionnaire using the same 9-point rating scale used previously. A copy of the Opinion Questionnaire and a table listing the means and standard deviations of each item on both administrations (as well as the means reported to the subjects on the second administration) has been deposited with the American Documentation Institute.[2]

Predictor Inventory. Jackson (1964) culled from two previous studies 45 true-false items which had successfully predicted social conformity. The first 22 items were taken from Barron (1953). Of 84 items rationally appearing to measure attributes related to experimental conformity, these 22 actually differentiated "yielders" from "independents." Barron described his procedure as follows:

> Most of these items were written anew, but others were culled from such sources as Murray's "Explorations in Personality," the E, F, and PeC scales of the California Public Opinion Study, and scales developed at the Institute of Personality Assessment and Research to measure such variables as Originality and Personal Soundness [p. 293].

The last 23 items of the Predictor Inventory were those MMPI and CPI items which differentiated yielders from independents in a study by Crutchfield (1955). A copy of the Predictor Inventory has also been deposited (see Footnote 2).

In the present experiment the Predictor Inventory was administered with two different sets of instructions. Approximately half the group received the following instructions:

> This is a test of your ability to understand other persons. It tests how well you can predict the attitudes and values of those who live around you, specifically University of Oregon freshman **women.**
>
> The inventory consists of 45 statements. Read each statement and decide whether others would consider a true or false answer characteristic of themselves. You are *not* asked whether the statement is true or false as applied to you; rather you are asked to decide which answer you think most other *University of Oregon freshman women* would endorse.

These instructions will be referred to as *prediction* instructions. The second group of subjects was asked to respond to each item of the Predictor Inventory as it applied to themselves. These traditional instructions will be referred to as *endorsement* instructions.

Procedure

The subjects in this study were tested 1 night each week, for approximately 1 hour each evening, for 5 consecutive weeks. They completed the first administration of the Opinion Questionnaire during the first testing session and the second administration during the fifth session. During each of the first four sessions each subject received either the endorsement or the prediction instructions for the Predictor Inventory. Half of the endorsement and half of the prediction group received the second

[2] This additional material may be obtained without charge from the authors at Oregon Research Institute, P. O. Box 5173, Eugene, Oregon 97403, or from the American Documentation Institute. Order Document No. 8564 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.75 for microfilm or $2.50 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

administration of the Opinion Questionnaire under memory instructions; the other half of each group received current instructions. Consequently, four experimental groups can be compared: endorsement-memory ($N = 34$), prediction-memory ($N = 35$), endorsement-current ($N = 36$), and prediction-current ($N = 52$).

RESULTS

Criterion Indices

From the double administration of the Opinion Questionnaire, two yielding indices were calculated for each subject. These criterion indices, described in relation to the means reported at Administration II, were: *yielding* (Y): The subject's average distance from the mean at Administration I minus her average distance from the mean at Administration II; and *yielding, corrected for distance from the mean* (Y/D): Yielding, divided by the subject's distance from the mean at Administration I. This latter statistic estimates the extent to which the subject yielded in relation to her opportunity to yield. An individual whose response on the first administration fell at the mean obviously would be unable to obtain a positive yielding score. On the other hand, an individual whose initial response deviated widely from the mean response might be expected to show some tendency to regress towards the mean. This corrected yielding statistic attempted to equate subjects in terms of their opportunity to yield. In the terminology of Fisher and Lubin (1958), Y is a measure of movement, while Y/D is a measure of conformity.

Each statistic was computed for the 25 items for which spurious mean values were reported, for the 20 items for which correct means were reported, and for all 45 items. The means, standard deviations, reliability coefficients, and intercorrelations of the criterion indices are presented in Table 1 for both the memory and current groups. In each case the figures in Table 1 were obtained by averaging the results for the prediction and endorsement groups. The corrected split-half reliability coefficients for the combined groups are listed in the diagonals. The column and row labeled r_D present the correlations between the yielding measures and the initial distance, D, of the subject from the reported mean.

As the diagonal entries in Table 1 indicate, the corrected split-half reliability estimates for the various yielding indices were substantial, though considerably less than unity. In general, Y/D had slightly higher reliability estimates than Y. The highest reliability coefficients were for Y/D over all 45 items, and the lowest were for Y over the 20 items with

TABLE 1

INTERCORRELATIONS, MEANS, STANDARD DEVIATIONS, AND RELIABILITY COEFFICIENTS FOR SIX CRITERION INDICES FROM THE OPINION QUESTIONNAIRE

Group	Memory groups ($N = 69$)						Current groups ($N = 88$)		
	Y			Y/D			M	σ	r_D
	Spurious	Correct	Total	Spurious	Correct	Total			
Current									
Y									
Spurious	.73[a]	.55	.93	.97	.60	.92	.98	.70	.18*
Correct	.48	.60[a]	.82	.57	.96	.79	.63	.49	.13
Total	.94	.77	.75[a]	.92	.83	.98	.82	.53	.19*
Y/D									
Spurious	.97	.49	.92	.79[a]	.62	.95	.29	.19	.01
Correct	.53	.96	.78	.56	.63[a]	.84	.27	.20	−.05
Total	.91	.74	.97	.94	.80	.82[a]	.28	.18	−.02
Memory									
M	.93	.72	.83	.27	.30	.28			
σ	.63	.52	.52	.19	.20	.18			
r_D	.05	.10	.08	−.16	−.03	−.09			

Note.—All correlations between indices are significantly greater than 0 ($p < .01$).
[a] Mean split-half reliability coefficient corrected by Spearman-Brown formula.
* $p < .05$; no other r_D correlations are significant.

correct means. The reliability coefficients based on the items given distorted means were slightly higher than those based on items given the correct means, but the reliability coefficients appeared to be unaffected by differences in instructions.

Because there were, by definition, more individuals close to the correct means than to the spurious means, the average amount of yielding (Y) to items for which the correct means were reported was, as expected, less than that to the items for which spurious means were listed. However, the Y/D values, which take this distance into account, were virtually identical for the two sets of items. The mean values of both Y and Y/D were similar for the current and memory groups, thus corroborating Hastorf and Piper's (1951) finding that directions to try to remember previous responses do not appreciably affect the total amount of yielding. The mean scores show that in all groups the subjects did tend to move in the direction of greater conformity with group opinion; in fact, they yielded from .67 to .99 point on a 9-point scale, or, as shown by the Y/D values, about 30% of the distance that they could have yielded if they had changed all of their answers to conform to the reported means exactly.

As Table 1 indicates, the intercorrelations among the yielding indices were all significantly different from 0 ($p < .01$). The correlations between the indices based upon spurious means and those based upon correct means were .55 and .48 for Y and .62 and .56 for Y/D; when these coefficients were corrected for attenuation (using the split-half estimates in Table 1), they were all in the .90s. Overall, the correlations between Y and Y/D ranged from .94 to .98, the virtual identity of these two indices most likely stemming from the fact that the subjects did not differ to any great extent in their initial distances from the composite reported means. The r_D correlations, ranging from −.05 to .19, also reflect the lack of individual differences in initial distance.

In summary, then, this instrument did pick up a marked tendency toward group conformity, and it measured yielding with respectable reliability. There were no significant differences in yielding as a function of either memory as opposed to current instructions, or accurate as opposed to inaccurate information.

Predictor Inventory

For each subject, the following scales (all keyed for independence) were scored for each administration of the inventory:

(All) All items from the Predictor Inventory ($N = 45$)
(B) All items from Barron ($N = 22$)
(C) All items from Crutchfield ($N = 23$)
(J) Jackson replication items ($N = 11$). These items are those 11 items from Jackson (1964) which differentiated yielders from independents when administered under desirability rating procedures. They can be compared with the following four sets of 11 items which differentiated similar groups under the more traditional endorsement procedures.
(Ba) Barron half-scale a ($N = 11$)
(Bb) Barron half-scale b ($N = 11$)
(Ca) Crutchfield half-scale a ($N = 11$)
(Cb) Crutchfield half-scale b ($N = 11$)

Scores on the following eight subscales allowed analyses of the effects of item keying:

(T) All true items ($N = 14$)
(Bt) Barron true items ($N = 8$)
(Ct) Crutchfield true items ($N = 6$)
(Jt) Jackson true items ($N = 5$)
(F) All false items ($N = 31$)
(Bf) Barron false items ($N = 14$)
(Cf) Crutchfield false items ($N = 17$)
(Jf) Jackson false items ($N = 6$)

The means, standard deviations, reliability coefficients, and intercorrelations among these scales for Administration I have been deposited (see Footnote 2). The average correlation between the Barron and Crutchfield scales was substantial (.53) for the endorsement groups, but nonsignificant for the prediction groups (.10). Most of the true and false scales were not interrelated for any of the four groups; at times, in fact, they correlated negatively. In marked contrast, the odd-even split-half reliability coefficients averaged .66 for the endorsement groups and .46 for the prediction groups. This wide range of coefficients for the three full-scale splits (Barron-Crutchfield, true-false, and odd-even) provides a vivid illustration of the effect item selection can have on split-half reliability estimates.

The split-half reliability coefficients for B and C averaged .45 for all four groups. For

the endorsement groups J appeared to be as reliable as B and C, but for the prediction groups the reliability of J was near 0. In general, split-half reliability estimates were higher for the endorsement groups than for the prediction groups. Moreover, the standard deviations of the scales were slightly lower for the prediction groups than for the endorsement groups, indicating some slight tendency for the subjects to share a common stereotype. In addition, the mean values for most of the scales were slightly lower for the prediction groups than for the endorsement groups, suggesting that the average subject may have ascribed a bit more originality and independence to herself than to her peers.

Since the Predictor Inventory was administered four times to each group, separate test-retest reliability estimates were calculated for Sessions 1 versus 2 and for Sessions 3 versus 4. The test-retest reliability coefficients for Sessions 3 versus 4 were higher than those for Sessions 1 versus 2 for all scales, confirming Howard's hypothesis that individuals become more consistent over time. Of equal interest is the fact that the reliability coefficients for the endorsement groups were substantially higher than the corresponding coefficients for the prediction groups, appearing to indicate that subjects were more consistent in their estimates of their own attributes than they were in their estimates of the attributes of others. For the total inventory (All), the test-retest correlations for the endorsement groups averaged .89 and .95, for Administrations 1 versus 2 and 3 versus 4, respectively; for the prediction groups the corresponding reliability coefficients averaged .72 and .84.

To test Howard's hypothesis that subjects move toward uniqueness over repeated testings, indices of interindividual similarity were computed for each administration of the inventory. Each subject's responses were compared with those of each of the other subjects in her group, and the average number of common responses was calculated. These indices, in turn, were then averaged for each group and for each administration of the inventory. The findings indicated that for this set of items the subjects responded in a fairly differential manner to begin with, and no evidence of increased interindividual variability occurred as a result of repeated testings. The average subject in the endorsement groups gave only 57% of her responses in common with others in her group on both the first and fourth administrations; for the prediction groups the corresponding figures were 61% and 60% for first and last administrations, respectively, again indicating some slight tendency for more stereotopy among the prediction groups.

TABLE 2

VALIDITY COEFFICIENTS FOR FIRST ADMINISTRATION

Scale	Endorsement groups		Prediction groups	
	Memory (N = 34)	Current (N = 36)	Memory (N = 35)	Current (N = 52)
All	−.06	−.17	−.30*	.03
B	−.28*	−.28*	−.39**	.07
C	.17	−.06	−.08	−.03
J	−.34*	−.19	−.13	−.16
Ba	−.20	−.27	−.30*	.11
Bb	−.29*	−.20	−.35*	.01
Ca	.33*	−.04	.18	−.07
Cb	−.07	−.09	−.37*	.03
T	−.31*	.07	−.16	−.04
Bt	−.31*	−.08	−.26	−.05
Ct	−.21	.18	.07	.00
Jt	−.46**	.11	−.03	−.21
F	.08	−.22	−.25	.05
Bf	−.19	−.28*	−.32*	.12
Cf	.28*	−.13	−.11	−.03
Jf	−.11	−.29*	−.15	−.06

Note.—Criterion = Y, based on 25 items with spurious means.
* $p < .05$.
** $p < .01$.

Validity Coefficients as a Function of Different Response Modes

Table 2 presents, for the first administration only, the correlations between the 16 Predictor Inventory scales and the yielding criterion, Y, based on the 25 items with spurious means. Validity coefficients for Y/D, essentially the same as those for Y, have been deposited (see Footnote 2). Table 2 indicates that the inventory as a whole had little validity, and *that* only for the prediction-memory group. (Note that, under the hypotheses of the study, valid coefficients are negative.) For the endorsement-memory group, which in the case of the total inventory might be considered a direct cross-validation, the correlation coefficients were nonsignificant. For this

group the Barron items carried all of the valid weight, with validity coefficients around −.30, while the Crutchfield items had positive coefficients; combining both sets of items destroyed the predictive validity of the inventory as a whole. For the prediction-memory group the Barron items were even more valid (−.39), and while the Crutchfield items were not significantly related to the criterion, they were in the predicted direction; consequently, the validity coefficients for the entire inventory were significant for this group. More of the significant validity coefficients were found among the two memory groups than among the two current groups.

Previous investigators (Jackson, 1964; Loomis & Spilka, 1963) have used the same items, keyed in the same direction, to compare the endorsement and desirability rating procedures. However, since the items in the Predictor Inventory were initially selected because of their predictive validity under endorsement conditions, there is every reason to suspect that such an analysis would favor the endorsement groups. To test adequately the major hypothesis of this study it is necessary to cross-validate two sets of items, each set of which previously discriminated yielders from independents when it was originally administered under its own set of instructions. Fortunately, Jackson has reported 11 items which differentiated at the .05 level or better yielders from independents under desirability rating instructions. For this set of 11 items (J) the present study provides a cross-validation. In order to provide comparison sets of items which originally differentiated under endorsement procedures, the Crutchfield and Barron items were subdivided into two odd-even half-scales, each composed of 11 items. Consequently, it was possible to compare the cross-validative results for the Jackson items under prediction conditions with the results for the four half-scales from Barron and Crutchfield under endorsement conditions. These data are also presented in Table 2.

The Jackson items, which originally differentiated under instructions similar to those given the prediction groups, not only failed to differentiate under these conditions on cross-validation, but were even more valid than the Barron items under endorsement procedures! In the cross-validation of the Jackson items afforded by the prediction-memory group, three of the four subscales from Barron and Crutchfield performed more validly. These results are the exact opposite of those expected.

The true subscales appeared to carry most of the valid weight under endorsement instructions, a finding congruent with that previously reported by Jackson. The false items, under endorsement instructions, appeared worthless. Under prediction instructions, on the other hand, both true and false subscales appeared about equally valid. It should be borne in mind that, since independence-conformity as measured by this inventory is a bipolar trait, the keying direction is arbitrary. Therefore, if the scale were keyed for conformity, it would be false items that were valid and true items that were invalid. The validity coefficients for each item for the first and fourth administrations of the Predictor Inventory have been deposited (see Footnote 2).

Effects of Repeated Testings

Table 3 presents the predictor-criterion correlations for four administrations of the Predictor Inventory. The results, reported only for the correlations with Y based on the 25 items with spurious means, show that, in general, repeated testings did not increase the validity coefficients. In fact, for the prediction-memory group, the originally significant validity coefficients vanished as a function of repeated testings. Validity coefficients for Y/D, essentially the same as those for Y, have been deposited (see Footnote 2).

In the last column for each group in Table 3 the validity coefficients for the mean scores over four administrations of the Predictor Inventory are listed. When these coefficients are compared with those from single administrations, the differences are slight. Since the hypothesis that the validity coefficients would tend to increase with repeated administrations was not confirmed, there would seem to be no reason for rejecting the mean score—or for that matter, the first score—as the best predictor.

TABLE 3

COMPARISON OF VALIDITY COEFFICIENTS OVER REPEATED TESTINGS

Scale	Endorsement groups					Prediction groups				
	Administration					Administration				
	1	2	3	4	*M*	1	2	3	4	*M*
Memory groups										
All	−.06	−.26	−.19	−.20*	−.20*	−.30*	.05	−.10	.04	−.08
B	−.28*	−.33*	−.29*	−.31*	−.32*	−.39**	.19	−.07	.05	−.05
C	.17	−.07	.00	−.07	.01	−.08	−.10	−.09	.00	−.07
J	−.34*	−.35*	−.35*	−.33*	−.37*	−.13	.09	.07	.04	.03
T	−.31*	−.42**	−.43**	−.49**	−.44**	−.16	−.08	−.26	−.09	−.17
F	.08	−.11	−.03	−.04	−.02	−.25	.08	.01	.08	−.01
Current groups										
All	−.17	−.20	−.22	−.19	−.20	.03	.08	−.03	.01	.02
B	−.28*	−.30*	−.34*	−.32*	−.33*	.07	.07	.09	.10	.09
C	−.06	−.07	−.05	−.06	−.06	−.03	.05	−.15	−.09	−.07
J	−.19	−.16	−.17	−.17	−.19	−.16	.05	−.08	−.04	−.06
T	.07	−.04	−.03	−.01	.00	−.04	−.03	−.19	−.04	−.09
F	−.22	−.22	−.25	−.23	−.24	.05	.10	.04	.03	.06

Note.—Criterion = *Y*, based on 25 items with spurious means.
* $p < .05$.
** $p < .01$.

DISCUSSION

The findings from this study show conclusively that knowledge of the responses of others can have a substantial effect upon subjects' responses, and that this effect occurs whether the information presented is accurate or inaccurate. This finding replicates similar findings by Hastorf and Piper (1951), Asch (1952), Barron (1953), Crutchfield (1955), Tuddenham (1958b), Jackson (1964), and Loomis and Spilka (1963).

The measures of yielding used in this study appear to be of sufficient reliability to recommend them for further experimental work. In particular, the relationship between yielding on the Opinion Questionnaire and submissive, acquiescent, or conforming behavior in other situations is certainly open to question, and some exploration of this relationship is clearly needed.

TABLE 4

VALIDITY COEFFICIENTS FROM THREE STUDIES

Study	Instructions	
	Endorsement	Prediction[a]
Jackson (1964)[b]	−.22*	−.29**
Loomis and Spilka (1963)	−.26*	.15
Goldberg and Rorer (present study)	−.06	−.30*

Note.—For the total 45-item Predictor Inventory (All) against a criterion of yielding (*Y*) for the 25 Opinion Questionnaire items for which spurious means were reported, with memory instructions.
[a] Social desirability rating instructions: Jackson (1964) and Loomis and Spilka (1963).
[b] Shifts away from reported means given value of 0 in calculating *Y*.
* $p < .05$.
** $p < .01$.

A comparison of the present study with previous ones reveals some puzzling differences. Table 4 presents the validity coefficients from the studies by Jackson (1964) and Loomis and Spilka (1963), along with those obtained from the present study when the entire 45-item Predictor Inventory (All) was correlated with yielding (*Y*) (based on the 25 questionnaire items with spurious means, administered under memory instructions). There were some procedural differences among the three studies. Both previous investigators obtained 9-point desirability ratings on the Predictor Inventory, rather than predictions of the endorsements of others as in the present study. Moreover, Jackson gave 0 value to shifts away from the reported means in computing *Y*. And finally, Jackson used a mixed group of male and female subjects, while the present study included only females (Loomis and Spilka did not report the percentage of males and females in their

study). However, these differences do not appear to explain the pattern of correlations reported in Table 4.

When these results are considered in conjunction with the paradoxical results obtained when the endorsement and the prediction scales were cross-validated within the present study, it seems safe to conclude that, to date, attempts to predict conformity behavior have been far from satisfactory. Furthermore, the disappointing validity results make it impossible to make any conclusive statements about the relative validity of either endorsement as opposed to prediction instructions, or single as opposed to repeated administrations of a test. In this latter regard, it should be noted that Howard's hypothesis of increased intra-individual consistency over repeated administrations was confirmed; however, Howard's hypothesis of increased interindividual variability was not confirmed, probably because of the initial low level of commonality for this set of items. The derivative hypothesis that validity should increase over repeated administrations may have been unsupported solely because the instrument had virtually no initial validity.

REFERENCES

ASCH, S. E. *Social psychology.* New York: Prentice-Hall, 1952.

BARRON, F. Some personality correlates of independence of judgment. *Journal of Personality,* 1953, 21, 287–297.

CAMPBELL, D. T. The indirect assessment of social attitudes. *Psychological Bulletin,* 1950, **47**, 15–38.

CRUTCHFIELD, R. S. Conformity and character. *American Psychologist,* 1955, **10**, 191–198.

DICKEN, C. F. Simulated patterns on the Edwards Personal Preference Schedule. *Journal of Applied Psychology,* 1959, **43**, 372–378.

EDWARDS, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden Press, 1957.

FISHER, S., & LUBIN, A. Distance as a determinant of influence in a two-person serial interaction situation. *Journal of Abnormal and Social Psychology,* 1958, **56**, 230–238.

FISKE, D. W. The constraints on intra-individual variability in test responses. *Educational and Psychological Measurement,* 1957, **17**, 317–337. (a)

FISKE, D. W. An intensive study of variability scores. *Educational and Psychological Measurement,* 1957, **17**, 453–465. (b)

GOLDBERG, L. R. The problem of "faking" in personality assessment. *Oregon Research Institute Technical Report,* 1962, **2**, No. 2.

GOLDBERG, L. R. A model of item ambiguity in personality assessment. *Educational and Psychological Measurement,* 1963, **23**, 467–492.

GORDON, L. V. The evaluation of personality by population judgments. *Journal of Social Psychology,* 1949, **30**, 305–309.

HASTORF, A. H., & PIPER, G. W. A note on the effect of explicit instructions on prestige suggestion. *Journal of Social Psychology,* 1951, **33**, 289–293.

HEILBRUN, A. B., JR. Social desirability and the relative validities of achievement scales. *Journal of Consulting Psychology,* 1962, **26**, 383–386.

HOWARD, K. I. Differentiation of individuals as a function of repeated testing. *Educational and Psychological Measurement,* 1964, **24**, 875–894.

HOWARD, K. I., & DIESENHAUS, H. I. 16 P.F. item response patterns as a function of repeated testing. *Educational and Psychological Measurement,* 1965, **25**, in press.

JACKSON, D. N. Assessing conformity with desirability judgments. *American Psychologist,* 1961, **16**, 446. (Abstract)

JACKSON, D. N. Desirability judgments as a method of personality assessment. *Educational and Psychological Measurement,* 1964, **24**, 223–238.

LOOMIS, R., & SPILKA, B. Social desirability and conformity in a test situation. *American Psychologist,* 1963, **18**, 358. (Abstract)

MESSICK, S. Dimensions of social desirability. *Journal of Consulting Psychology,* 1960, **24**, 279–287.

MITRA, S. K., & FISKE, D. W. Intra-individual variability as related to test score and item. *Educational and Psychological Measurement,* 1956, **16**, 3–12.

MOORE, H. T. The comparative influence of majority and expert opinion. *American Journal of Psychology,* 1921, **32**, 16–20.

MÜNSTERBERG, H. *Grundzüge der Psychotechnik.* Leipzig: Barth, 1914.

SCOTT, W. A. Social desirability and individual conceptions of the desirable. *Journal of Abnormal and Social Psychology,* 1963, **67**, 574–585.

TUDDENHAM, R. D. The influence of a distorted group norm upon individual judgment. *Journal of Psychology,* 1958, **46**, 227–241. (a)

TUDDENHAM, R. D. Studies in conformity and yielding: VI. The influence upon judgment of a genuine norm. Technical Report No. 7, 1958, University of California, Contract No. NR 170-159, Office of Naval Research. (b)

(Received July 23, 1964)