XEROX

Group Psychology Branch
Office of Naval Research

Research Performed Under Contract Nonr 1288 (04)

Third Annual Report

VARIABLES RELATED TO ACCURACY IN INTERPERSONAL PERCEPTION

by

Victor B. Cline

and

James M. Richards, Jr.

Department of Psychology
University of Utah
Salt Lake City
Utah

November 15, 1960

THIRD ANNUAL REPORT

Variables Related to Accuracy in Interpersonal Perception

Victor B. Cline and James M. Richards, Jr.

University of Utah

INTRODUCTION

For the past three years research has been conducted in the area of person perception at the University of Utah with grant support from the Group Psychology Branch of the Office of Naval Research. The basic approach has been to use sound-color movies of people in interview situations as stimulus material for judges in making evaluations and judgments about other people. In the three years that this project has been under way the number of films used has been reduced from 24 to 6, using methods approximating standard item analysis. Heterogeneous groups of judges have viewed these sound movies and made a variety of predictions about the real life, verbal, and test behavior of the subjects seen in the films. One of the major early findings was that there exists a certain amount of generality in ability to judge accurately across judging instruments, as well as across films. This finding has held up in some seven independent experiments reported in detail in the first two Annual Reports (Cline and Richards, 1958, 1959).

Cronbach (1955) has, in the past, strongly criticized global approaches to obtaining accuracy of judgment scores. He feels that global scores are virtually meaningless because they lump together all sorts of response sets, and to meet this problem he has proposed several complex methods of securing so-called "component scores." In previous research, analysis of some of these component scores suggested by

Cronbach led to the conclusion that even though there is a modest degree of generality in judging ability there are also at least two major types of judging accuracy ability. The first major type of ability was "Sensitivity to the Generalized Other" (or, in Cronbach's terminology, "Stereotype Accuracy"), and the second major type was "Interpersonal Sensitivity" or, in Cronbach's terms, "Differential Accuracy" and/or "Differential Elevation." These results were obtained using the Trait Rating type of instrument where judges rated the persons interviewed using a Likert type six category scale. Bronfenbrenner and his associates (1958) in a completely independent study obtained essentially the same results.

However, difficulties were repeatedly encountered in working with and interpreting some of Cronbach's components of accuracy scores (using the Trait Rating instrument). In going back over Cronbach's article it was found that he appeared to contradict himself in discussing the meaning and interpretation of some of these components. In addition, the conclusion was reached that making ratings on rather haphazardly chosen traits is not a very meaningful judging task, and it was therefore decided that a new judging instrument was needed which would allow clarification of the use and interpretation of these component accuracy scores. Since part of the filmed interview which the judges saw involved questions relating to one's religious values and beliefs, a new instrument called the Belief-Values Inventory was developed. On this instrument, the judge is required to predict the responses of the person seen in the film to such questions as, "The idea of God is just a fiction," on a five category Likert type scale.

This proved to be a much more satisfactory instrument than the

Trait Rating measure, and accordingly in the past year a major study of Belief-Values Inventory component scores was carried out. This study, which is reported in detail later in this report (see page), has suggested some modifications in Cronbach's analytic scheme.

BRIEF RESUME OF THIRD YEAR'S WORK

The third year (1959-1960) emphasized research in the following areas:

1. An intensive study of the analysis of accuracy scores into components, using an analytic scheme similar to, but with some modifications of, the scheme suggested by Cronbach (1955).
2. Analyses of the data of an experiment in which a comparison of the judgments of individuals versus groups (where group members collaborate in decision and judgment making) is made.
3. An investigation was made of the clinical versus statistical prediction problem in terms of components of accuracy scores.
4. Sponsoring a group discussion type symposium entitled New Frontiers in Person Perception Research at the annual meetings of the American Psychological Association, September, 1960, in Chicago, Illinois.
5. The development of new interviews (to be filmed) was undertaken and is currently in progress.
6. The further development and refinement of new judging instruments was initiated on the basis of earlier research findings with components of accuracy scores. This work also is in progress.

CHAPTER I

Accuracy Components in Person Perception Scores and the Scoring System as an Artifact in Investigations of the Generality of Judging Ability

The usual procedure in investigations of accuracy of person perception is to have a judge predict how some other person has responded to, or will respond to, some standard instrument. Then the judge's predictions are compared to the other person's actual responses, and some measure of the degree of agreement between the two is taken as a measure of the judges' accuracy of person perception.

A special case of this procedure is that in which the judges make their predictions and the "others" make their responses on a standard instrument which consists of a series of items, each of which has several possible responses differing along a scale to which numerical values can be applied. An example of such a scale would be a series of adjectives each of which was rated by the "other" for the degree to which it was descriptive of himself on a six category scale ranging from "Very Like" to "Very Unlike," and with corresponding numerical scores assigned ranging from 1 for "Very Like" to 6 for "Very Unlike." The measure of accuracy of person perception used with this type of instrument is based on the numerical difference between the judge's predictions and the "other's" actual responses. The usual procedure in investigations of this type is to square this difference score for each item and then average across items. This average value is commonly called the "D^2 statistic." (Cronbach and Glaser, 1953)

Cronbach (1955) has strongly criticized treatment of accuracy of person perception in this global fashion, and has proposed a more analytic treatment based on mathematically independent components of the D^2 statistic. Cronbach's analysis is most clearly pertinent to those situations in which judges predict the responses of a series of "others," rather than a single "other." In this case, Cronbach treats these judgments as a matrix $\underline{X}$ in which the "others" whose responses are predicted compose the columns of the matrix, and the items on which predictions are made compose the rows of the matrix. Thus, each element, x_{ij}, of this matrix indicates that "other" number j was predicted to have made the response with numberical value $\underline{x}$ on item number $\underline{i}$. There is one such matrix for each judge. In order to compute the D^2 measure of accuracy of prediction and its various components, each such matrix is compared to a criterion matrix $\underline{C}$, in which each element c_{ij} is the numberical value of the actual response of "other" $\underline{j}$ on item $\underline{i}$. In Cronbach's scheme, D^2 is broken down into four basic components based on discrepancies between numerical values of predicted responses and numerical values of actual responses. Three of these four components can be further broken down into a correlation term and a variance term. These components are:

(1) Elevation - This is a measure of the difference between the mean of the numerical values in the judgment matrix taken over all "others" and all items and the corresponding mean of the numerical values in the criterion matrix. Cronbach states that this component is primarily a measure of the way in which each judge uses the rating scale rather than a measure of judging ability.

(2) Differential Elevation - This component is a measure of the difference between the means of the columns of the judgment matrix and

the means of the columns of the criterion matrix, with the contribution of the elevation component eliminated. It is thus a measure of the ability to predict differences between "others" taken across items. This component can be further broken down into (a) a correlation term which measures the extent to which the judge arranges "others" in the same order in which they are ordered in the criterion matrix, and (b) a variance term which measures the extent to which judged differences between "others" are large or small.

(3) Stereotype Accuracy - This component is a measure of the difference between the means of the rows (or items) of the judgment matrix and the corresponding means of the criterion matrix, again, with the contribution of the elevation component removed. This component also can be broken down into (a) a correlation term and (b) a variance term, and is a measure of how accurately the judge predicts the responses of the typical "other" and/or people-in-general to the items.

(4) Differential Accuracy - This component is a measure of difference between the scores for "others" on individual items in the judgment matrix and the corresponding scores in the criterion matrix, where in each case an individual "other's" score is taken as a deviation both from his own mean and from the item mean. This component can also be broken down into a correlation term and a variance term, and is averaged across items to obtain an overall measure. Cronbach considers this component and particularly its correlational term the most appropriate measure of what is ordinarily meant by accuracy of person perception.

Recent research by Bronfenbrenner and his associates (1958) and by the authors (Cline and Richards, 1958) has suggested that such an analytic scheme is a promising approach in studies of person perception.

Bronfenbrenner, using a similar, but not identical, approach, found two relatively independent aspects of ability to predict the responses of others: first, "Sensitivity to the generalized other" (a measure similar to Cronbach's Stereotype Accuracy) which is the ability to predict the mean responses of several people on individual items, and "Individual sensitivity," (a measure similar to Cronbach's Differential Accuracy) which is the ability to order correctly individuals on items. In the research of the authors, Cronbach's scheme, in its difference score form, was used specifically. The judging instrument used in this study was the Trait Rating test, a list of twenty-five adjectives which judges rated on their degree of similarity to ten standard "others" presented by means of a sound color movie. The results of this experiment suggested that over-all judging ability consisted mainly of two independent parts: "Stereotype Accuracy" and "Differential Accuracy" (using Cronbach's terminology).

The results of these studies appear at first glance to offer striking confirmation of each other, particularly so since quite different procedures were used; and in combination to confirm the importance of treating person perception scores in the way suggested by Cronbach. In spite of the promise of these results, however, in subsequent research the authors encountered persistent difficulties in the interpretation of these components, particularly when an attempt was made to break down Differential Elevation, Stereotype Accuracy, and Differential Accuracy into a correlation term and a variance term. This difficulty came, through the course of much work, to be focused on the Differential Elevation component, and its correlation term, largely because Cronbach (1955) appeared to contradict himself about the interpretation of these measures on the same page, first stating that

these measures reflect primarily whether or not the judge interprets the words defining the scale in the same way as the "others" do, and that therefore they appear "relatively unfruitful" as a source of information on his perception of "others," and second, stating that these scores are measures of the judge's sensitivity to individual differences.

In attempting to clear up these difficulties of interpretation, the authors felt they needed an instrument in which the predictions themselves could be interpreted more easily than they could using a typical Trait Rating test. They therefore developed a new judging instrument, the Belief-Values Inventory. This judging instrument required the judge to predict "others" responses to twelve Likert type items dealing with religious beliefs and values. A sample item is:

> When in doubt, I have found it best to stop and ask God for guidance.
>
> A. strongly agree
> B. agree
> C. neither agree nor disagree
> D. disagree
> E. strongly disagree

In the filmed interview the "other" had been asked direct questions about his attitudes toward religion.

Subsequent research with this instrument indicated that the difficulties with this type of analytical treatment arise largely because the scoring system has a differential effect upon the components involved in a Cronbach type analysis. With items of the type included in the Belief-Values Inventory, there are two possible scoring methods. The first of these is to score "Strongly Agree" as 1, "Agree" as 2, etc., without regard to whether or not on that particular item, "Strongly Agree" is a pro-religious answer. The second possible scoring system is to score the most conventional pro-religious response

as 1, regardless of whether that answer is "strongly Agree" or "Strongly Disagree." If the first of these scoring systems is used, the Stereotype Accuracy variance is large, but the Differential Elevation variance is made artificially small. On the other hand, if the second of these scoring systems is used, the Stereotype Accuracy variance is artifically reduced, while the Differential Elevation variance is maximized. These effects are illustrated by Table I, which presents the responses of three hypothetical persons to items of this type, with each item score presented in both of the two scoring systems. In Table I, the first hypothetical person always answered with the most conventional religious answer; the second hypothetical person always answered with the second most conventional religious answer, and the third hypothetical person always gave the middle or neutral response. In this table the consistency of responding by each person is, of course, somewhat exaggerated to make the point clear.

In addition to the effects of the scoring system on the variance of Stereotype Accuracy and Differential Elevation, several other things are apparent from Table I. The first of these is that if these three hypothetical persons were used as "others" in our films and the first scoring system were used (i.e., where "Strongly Agree" is always scored as 1, without regard to whether this is in the religious or non-religious direction) no matter what degree of accuracy a judge attained in predicting their responses the Differential Elevation correlation component could take no other value than .00. If one then tried to relate the Differential Elevation correlation term to other measures of judging values, it is obvious that there could be no relationship, and one might erroneously conclude that there is no

Table I

Comparisons of Responses of Three Hypothetical Persons* to Belief-Values

Inventory Type Items Using Two Different Scoring Systems

Inventory Item	Scores of three hypothetical people (A, B, & C) when Strongly Agree is always scored 1.			Sterotype Accuracy $\bar{X}$
	A	B	C	
I believe in God:	1 (Strongly Agree)	2 (Agree)	3 (Neither	2
Religion is Nonsense:	5 (Strongly Disagree)	4 (Disagree)	3 Agree	4
Prayers are answered:	1 (Strongly Agree)	2 (Agree)	3 nor	2
All churches should be closed:	5 (Strongly Disagree)	4 (Disagree)	3 Disagree)	4
Differential Elevation $\bar{X}$	3	3	3	

	Scores of three hypothetical people (A, B, & C) when the most pro-religious answer is always scored 1.			Stereotype Accuracy $\bar{X}$
	A	B	C	
I believe in God:	1 (Strongly Agree)	2 (Agree)	3 (Neither	2
Religion is nonsense:	1 (Strongly Disagree)	2 (Disagree)	3 Agree	2
Prayers are answered:	1 (Strongly Agree)	2 (Agree)	3 nor	2
All churches should be closed:	1 (Strongly Disagree)	2 (Disagree)	3 Disagree)	2
Differential Elevation $\bar{X}$	1	2	3	

*In the above tables person A always responds in the most conventional pro-religious fashion. Person B responds in the second most pro-religious manner (i.e., "Agree" instead of "Strongly Agree", and "Disagree" instead of "Strongly Disagree" where appropriate). Person C always chooses the neutral middle category, "Neither Agree nor Disagree."

generality of judging ability. Use of the second scoring system (i.e., where the most pro-religious response is always scored 1 regardless of whether it is "Strongly Agree" or "Strongly Disagree"), however, has a similar effect on the Stereotype Accuracy correlation and again could lead to a false conclusion that there is no generality of judging ability. It will also be seen from Table I that if the first scoring system were used in a study of judging ability, Differential Elevation and also Elevation would reflect primarily the extent to which judges interpreted items in the same way as the "others," but that if the second scoring system were used Differential Elevation would be a measure of the judges' "sensitivity to individual differences" in overall religiosity, and Elevation would be a measure of the judged average religiosity of the group of "others." Thus the apparent paradox in Cronbach's formulation is resolved. All of this, taken together, strongly suggests that in investigation of accuracy of person perception, and particularly of its generality, neither of these scoring systems is by itself satisfactory, but rather that the first scoring system should be used in computing Stereotype Accuracy and its components and the second scoring system should be used in computing Differential Elevation and its components. It should be emphasized that, in the hypothetical example, the items differed greatly in their average degree of endorsement and the persons differed greatly in their over-all degree of religiosity. In studies of accuracy of interpersonal perception, both of these would be important and yet either one or the other would inevitably be artificially eliminated if either scoring system alone was used.

These two scoring systems also have another effect which is not readily apparent in the hypothetical example, since it eliminates all

the persons-by-items interaction that would occur in a real problem. This effect is that the Differential Accuracy component and its correlation and variance constituents will all take on different values depending on which scoring system is used, and the authors are aware of no criterion which would indicate in a real problem which of these values are the most appropriate measures of judging ability. This led the authors to the conclusion that none of the values of Differential Accuracy and its constituents are particularly good measures of judging ability, and to drop it from further consideration in their research. They have been replaced with a new index of judging ability which the authors have chosen to call Interpersonal Accuracy. There is no difference score form of this measure; it consists only of a correlation term and a variance term. The correlation term is computed by determining the correlation between each judge's predicted values and the corresponding actual responses by "others" on individual items and then averaging across items (without converting these scores in terms of their discrepancy from item and person means as is the case with Differential Accuracy). Similarly, the Interpersonal Accuracy variance term involves the computation of the variance of each judge's predictions on individual items, averaged across items. With regard to how this index fits into Cronbach's scheme, the authors are of the opinion that it is a linear combination of Differential Elevation and Differential Accuracy. It offers the strong advantage over other measures that it is invariant under changes of scoring system.

There are several additional considerations in the interpretation of this hypothetical example. The first of these is that there is a strong general "religiosity" factor underlying the questions. In

Cronbach's scheme, this general factor should be tapped by the Differential Elevation component, and this does occur under the scoring system in which the conventional religious answer is always scored 1. It might be objected, therefore, that the scoring system in which "Strongly Agree" is always scored 1 "randomizes" this general factor. This effect of this scoring system is, however, exactly the point of the hypothetical example; and it should be emphasized again that the scoring system in which "Strongly Agree" is always scored 1 is the only scoring system which permits the real differences in the average degree of endorsement of the items (Stereotype Accuracy) to appear. It should also be noted that the nature of the questions presented in the hypothetical example was intentionally made such as to emphasize the inappropriateness of this scoring system for the Differential Elevation component. In a less extreme, more realistic case, this inappropriateness would be less clear. The fact that there is a strong general factor does contribute greatly to the clarity of interpretation of the Interpersonal Accuracy component, and the present authors are in agreement with Cronbach's position that several factorially pure sets of items analyzed separately are preferable to one factorially complex set of items treated in a global fashion.

If the argument of the authors is accepted to this point, it is still an open question whether the considerations outlined have any practical effect on investigations of accuracy of person perception. A study providing some information with regard to this point has been conducted. In this experiment 46 undergraduates, both male and female, at the University of Utah, predicted the responses of six standard others, presented through the filmed interview procedure, on the

Belief-Values Inventory. Details of the experimental procedure of using these filmed interviews are presented elsewhere. (Cline and Richards, 1960 A) Using a program developed for the IBM 650 Computer, the predictions of these judges were scored twice against the criterion, once with each of the two scoring systems discussed above, and the various judgment scores intercorrelated. When this program is used all correlation terms are expressed in terms of Fisher's Z. Results of both of these analyses are presented in Table II. In Table II, correlations above the diagonal were obtained when "Strongly Agree" was always scored 1 regardless of whether or not it represented a pro-religious answer, and correlations below the diagonal were obtained when the most conventional pro-religious answer was always scored 1. On the basis of either of these two groups of correlations alone, one would have to conclude that there is no consistent pattern of generality in judging ability, particularly as reflected in the three correlation measures, but rather an appearance of two relatively independent factors measured respectively by the Differential Elevation correlation term and the Stereotype Accuracy correlation term, thus confirming the previous results of Bronfenbrenner and his associates (1958) and Cline and Richards (1960 A).

A further analysis of these data was made, however, in which judgment scores were intercorrelated across scoring systems in such a way that each component is scored most appropriately. More specifically, Stereotype Accuracy and its correlation and variance terms were computed using the scoring system where "Strongly Agree" was always scored 1, and Differential Elevation and its components and Elevation were computed using the scoring system where the most pro-religious answer was always

Intercorrelations of Belief-Values Inventory Components
for Each of Two Scoring Systems

	Total	Elevation	Differential Elevation	Stereotype Accuracy	Differential Elevation Z	Stereotype Accuracy Z	Interpersonal Accuracy Z	Differential Elevation σ^2	Stereotype Accuracy σ^2	Interpersonal Accuracy σ^2
Total		.00	.24	.37	.21	.72	.84	-.40	.46	.07
Elevation	.55		.03	-.11	-.06	-.07	.11	.03	-.16	.18
Differential Elevation	.38	.36		.13	.80	.06	.30	-.41	.08	-.20
Stereotype Accuracy	.01	.11	-.17		.06	.95	.56	-.15	.58	.35
Differential Elevation Z	.67	.46	.80	.01		.05	.25	.06	.08	-.03
Stereotype Accuracy Z	.19	.02	.05	.53	.18		.51	.10	.45	.32
Interpersonal Accuracy Z	.84	.62	.72	.08	.86	.23		-.16	.37	.33
Differential Elevation σ^2	.28	.42	-.07	.21	.33	.26	.25		.06	.47
Stereotype Accuracy σ^2	-.11	-.17	.22	-.71	.10	.17	.05	-.10		.40
Interpersonal Accuracy σ^2	.07	.33	-.21	.20	.22	.22	.33	.96	-.13	

r .05 = .29 r .01 = .37

Correlations above the diagonal were obtained when Strongly Agree was always scored 1, and correlations below the diagonal were obtained when the most conventional religious answer was always scored 1.

All Z components are Fisher's Z transformations of Pearson correlation.

All correlations between Total and difference score components are corrected part-whole correlations.

scored 1. Results are presented in Table III. In this table, there is a consistent pattern of a significant degree of generality across the correlation terms of all components, thus suggesting that judging ability is, to some degree, a general trait.

These results, in the opinion of the authors, clearly indicate that the scoring system may be an important artifact in investigations of the generality question when using components of accuracy scores, and therefore strongly supports the argument advanced in the hypothetical example discussed earlier. It is still most important to avoid over-generalization from these results. It is still possible, and even probable, that when using other judging instruments, other judges, or other persons to be judged, accuracy of stereotype and accuracy of judgments of individual differences may prove to be <u>really</u> independent. Future investigators should, however, on the basis of the results of this study, guard against a false conclusion that the two main types of accuracy are independent when that independence represents nothing more than scoring system artifact. The results also suggest that <u>Interpersonal Accuracy</u> is the most appropriate measure of the ability to judge accurately individual differences, at least in those cases where a strong general factor is present in the items on which the ratings are made.

Intercorrelations of Belief-Values Inventory Components
When Each Component is Scored Appropriately*

	Total	Elevation	Differential Elevation	Stereotype Accuracy	Differential Elevation Z	Stereotype Accuracy Z	Interpersonal Accuracy Z	Differential Elevation σ^2	Stereotype Accuracy σ^2	Interpersonal Accuracy σ^2
Total	—									
Elevation	.55	—								
Differential Elevation	.33	.36	—							
Stereotype Accuracy	.37	.93	.25	—						
Differential Elevation Z	.67	.46	.80	.40	—					
Stereotype Accuracy Z	.72	.82	.22	.95	.38	—				
Interpersonal Accuracy Z	.84	.62	.72	.56	.86	.51	—			
Differential Elevation σ^2	.28	.42	-.07	.48	.33	.43	.25	—		
Stereotype Accuracy σ^2	.46	.70	.18	.58	.26	.45	.37	.48	—	
Interpersonal Accuracy σ^2	.07	.33	-.21	.35	.22	.32	.33	.96	.40	—

r .05 = .29 r .01 = .37

*All Z components are Fisher's Z transformations of Pearson correlations.

All correlations between Total and difference score components are corrected part-whole correlations.

CHAPTER II

A Comparison of Individuals vs Groups in Judging Personality

As a practical necessity men are continually required to subjectively judge, assess, and evaluate their associates. Frequently in the military or in industry this is a prerequisite in initial employment, promotion, etc. There have been various approaches to the quantification of subjective judgments of which perhaps the most common have been rating procedures. Since this type of judgment and the decisions or courses of action which results therefrom have so many far reaching implications, any research which might further contribute to our knowledge in this area should be of considerable intrinsic importance. The purpose of the present study was to (1) determine whether individuals or groups are more likely to be accurate in making social judgments (i.e. "predictions" of the behavior and personality of other individuals), and (2) at the same time compare different types of group judgment. These judgments were made on instruments similar to two different kinds of rating scales commonly used in applied settings.

The rationale of this experiment grew out of the recent survey of studies comparing group performance and individual performance made by Lorge, Fox, Davitz, and Brenner (1958). The general conclusion of this survey was that a group, on almost any task, will perform better than a typical individual, but not necessarily better than a superior individual on the task in question. This finding is true whether the "group performance" is made by a genuine group or is merely a statistical combination of several independent individual performances. An

unresolved question is the degree to which these findings can be attributed to a reduction in the variability of the group performance.

The trend of the studies cited in this survey suggested the hypotheses to be tested in this experiment. These hypotheses are:

1. The accuracy of predictions (about the behavior of other persons) made by a group of persons arriving at a consensus prediction through group discussion will be significantly greater than the average accuracy of the predictions made by the individuals composing the group. The average accuracy of the predictions made by the individual composing the group will also be significantly less than the accuracy of an "artificial group" (i.e., a single prediction derived through a statistical combination of their individual predictions) and also less than the accuracy of prediction of the best individual among the individuals composing the group.

A secondary question has to do with the presence or absence of a consistent pattern of superiority in accuracy among predictions made by best individual judges, consensus groups, and "artificial groups."

Method

The subjects were 186 students, both male and female, in the introductory psychology classes at the University of Utah in the Fall of 1959. The procedure involved the presentation of six filmed interviews or "standard others." These were photographed in sound and color, and were conducted by an actor, a member of the University Theatre staff, who asked a fairly standard series of questions (to insure equivalence over interviews) probing the following areas: (a) personal values, (b) personality strengths and weaknesses, (c) reaction to the interview, (d) hobbies and activities, (e) self-conception, and (f) temper.

After a filmed interview had been shown the projector would be stopped and the subject-judge required to fill out paper-pencil judging instruments. Following this another interview would be shown and so forth. Details of the development and selection of these films, the experimental procedures involved, and certain underlying methodological and theoretical considerations have been published elsewhere (Cline & Richards, 1958, 1960 A).

In this study, two prediction instruments were used. The first of these was the Adjective Check List, which required the subject to determine which of a pair of adjectives the interviewee had checked as being descriptive of himself. A sample item is:

14. _____ (a) resourceful

_____ (b) cheerful

There were 20 such pairs for each of the six films making a total of 120. The score on the Adjective Check List was the number correct. Thus the Adjective Check List is similar to a forced-choice rating procedure.

The second instrument used was the Belief-Values Inventory. On this instrument the subject was required to determine (predict) how the interviewee had responded to a Likert type scale dealing with religious beliefs. During the course of the interview, the person in the film had been asked direct questions in this area. A sample item is:

I feel quite sure God does not exist.

_____ (1) Strongly agree

_____ (2) Agree

_____ (3) Neither agree nor disagree

_____ (4) Disagree

_____ (5) Strongly disagree

Thus the Belief-Values Inventory is comparable to a graphic rating procedure.

There were 12 such items for each film or interview. Several different scores based on a recent modification by Cline and Richards (see Chapter I) of an analytic procedure suggested by Cronbach (1955) were computed from judges' responses to this instrument using a program developed for the IBM 650 Computer. The first of these was a total score, which was based on the average of the squared discrepancies (using the one to five point scale) between predicted responses by each judge for each interviewee, and actual responses of each interviewee. This is an error score, and in order to make these scores comparable to other scores used in this study, the scores were converted to accuracy scores through a standard score transformation, setting the mean equal to 50 and standard deviation equal to 10.

The second two BVI scores are components of what Cronbach (1955) has called Stereotype Accuracy. This measures the degree to which each judge predicts how the group of interviewees as a whole responds to the judging instrument, and involves the degree to which the means of items (averaged across interviewees) predicted by each judge corresponds to actual item means. The two scores used in this study are the (1) correlation between each judge's predicted item means and obtained item means, converted to a Fisher's Z, and the (2) variance of each judge's predicted means. Cronbach has demonstrated these two scores to be the two parameters in Stereotype Accuracy when the criterion is held constant, and they permit independent evaluations of the effect of grouping on accuracy and on variability of prediction in this study.

The last two scores on the BVI are measures of Interpersonal

Accuracy. This represents the degree to which judges accurately predict the responses of interviewees to individual items, and involves mainly the degree to which judges correctly order the interviewees in terms of their overall degree of "religiosity." It, therefore, is the best measure of the kind of accuracy that is the main concern in most institutional rating situations. Interpersonal Accuracy, like Stereotype Accuracy, has two independent parameters, a correlation term expressed in terms of Fisher's Z, and a variance term, thus permitting independent evaluation of accuracy and variability. The correlation score is computed by determining the correlation between each judge's predicted values and the corresponding actual values on individual items, converting to Fisher's Z and averaging across items. The variance score is computed by determining the variance of each judge's predicted scores on individual items and averaging across items.

Procedure

The 186 subjects in this experiment were divided into 62 three-person groups. The division was made at the time the experiment was conducted, and most groups consisted of three persons seated next to each other in the experimental room. Group composition in terms of sex of group members was roughly random. The subjects saw each film and first completed the judging instruments independently. They then joined together in group discussion fashion and proceeded to arrive at a consensus judgment for the items on the judging instruments without referring back to, or looking at, their earlier independent judgments.

The "artificial group" judgment was derived from the first individual judgments of the group members. Thus, on the Adjective

Check List, the "artificial group" judgment was determined on the basis of a "majority vote" of the judges on each item (by inspecting their individual judging protocols). On the Belief-Values Inventory, it was calculated by determing the average of the values predicted by the three judges for each interviewee on each item. It is important to emphasize that this "artificial group" is not a group in the psychological sense, but only a statistical combination of the original independent judgments.

The "average accuracy of individuals composing the group" was, of course, obtained by computing the mean of the accuracy scores of the three individuals who made up each group. It is most important to note that this is not the same thing as the "artificial group" procedure where it was the actual predictions of the three group members that were averaged rather than their accuracy scores.

The "best judge" in each group was selected on the basis of his accuracy scores. In interpreting the results of this study, therefore, it is important to note that this selection was done on an after-the-fact basis, thus maximizing accuracy scores for this condition by capitalizing on chance. It would, therefore, be impossible for a "best judge" selected in advance to obtain a higher score than this, and such a "best judge" would, in fact, probably score somewhat lower, since some error would be involved in any advance selection. The best judges were selected independently for the ACL and the BVI and therefore were not necessarily the same person on the two different instruments. On the BVI, however, the "best judges" selected on the basis of total score, were also used as "best judges" in making the comparisons involving the other scores derived from this instrument.

Results

The means and standard deviations for each judgment procedure on each judgment score are presented in Table 1. In table 1, all scores are accuracy scores. Since total score on BVI is based on error score, in Table 1 this judgment score is transformed to a standard score distribution with mean = 50 and standard deviation = 10.

As a first step in the statistical analysis of these data, overall F tests were calculated for each of the judgment scores separately. The results of this analysis are presented in Table 2. No test for homogeneity of variance was made before calculation of these F tests. This procedure was followed because the recent work of Boneau (1960) strongly suggests that F is not significantly affected by heterogeneity of variance if the sample sizes are identical and relatively large, i.e., 20. Both of these conditions hold in the present study. It is also known that available tests for homogeneity of variance are affected too much by other variables than that involved in the null hypothesis to justify their use prior to an analysis of variance (Box, 1953).

Since all of the F tests in Table 2 are significant at or beyond the .01 level of confidence, a test for significance of difference between individual means was made. This test was made using the Multiple Range Test (Li, 1957, p. 238), which is the most appropriate procedure known to the experimenters for making "post-mortem" type comparisons between individual means after an overall F test has been made. Briefly, the Multiple Range Test involves computing a value which represents how large the difference between two means must be in order to be significant at a stated level, and then comparing the obtained difference to this value. Results of this analysis are summarized in Table 3.

Table 1

Means and Standard Deviations of Judgment Scores

	Average of Individuals Composing The Group	"Best Judge"	Group Consensus	Artificial Group
Adjective Check List				
$\bar{X}$	97.27	101.66	102.52	103.32
σ	3.51	3.91	3.95	4.55
Belief Values Inventory Total				
$\bar{X}$	43.29	53.92	49.47	52.87
σ	8.61	8.31	11.16	8.02
Belief Values Inventory Stereotype Accuracy Z				
$\bar{X}$	1.19	1.44	1.28	1.41
σ	.28	.37	.45	.37
Belief-Values Inventory Stereotype Accuracy Variance				
$\bar{X}$	.35	.40	.31	.30
σ	.14	.22	.15	.13
Belief-Values Inventory Interpersonal Accuracy Z				
$\bar{X}$	.90	1.01	1.00	.98
σ	.12	.14	.16	.15
Belief-Values Inventory Interpersonal Accuracy Variance				
$\bar{X}$	1.09	1.06	1.06	.91
σ	.23	.28	.31	.25

Table 2

Results of Overall F Tests
For Judgment Scores

Judgment Score	Between Variance d.f. = 3	Within Variance d.f. = 244	F	p
Adjective Check List Total	451.82	13.82	32.62	.001
Belief Values Inventory Total	1423.02	84.33	16.87	.001
Belief Values Inventory Stereotype Accuracy Z	.8633	.1432	6.03	.001
Belief Values Inventory Stereotype Accuracy Variance	.1333	.0282	4.72	.01
Belief Values Inventory Interpersonal Accuracy Z	.1633	.0213	7.67	.001
Belief Values Inventory Interpersonal Accuracy Variance	.3900	.0754	5.17	.01

Table 3

Tests for Significance of Difference Between Individual Means for Each Judgment Score

Judgment Score	Average of Ind. vs. Best Judge	Average of Ind. vs. Grp. Cons.	Average of Ind. vs. Art. Grp.	Best Judge vs. Grp. Consensus	Best Judge vs. Art. Grp.	Grp. Cons. vs. Art. Grp.
Adj. Check List Total Diff. Bet. Means	4.39**	5.25**	6.05**	.86	1.66*	.80
Belief Values Inv. Tot. Diff. Bet. Means	10.63**	6.18**	9.58**	4.45**	1.05	3.40*
Belief Values Inv. Stereo. Acc. Z Diff. Bet. Means	.25**	.09	.22**	.16*	.03	.13*
Belief Values Inv. Stereo. Acc. Var. Diff. Bet. Means	.05	.04	.05	.09**	.10**	.01
Belief Values Inv. Inter. Acc. Z Diff. Bet. Means	.11**	.10**	.08**	.01	.03	.02
Belief Values Inv. Inter. Acc. Var. Diff. Bet. Means	.03	.03	.18**	.00	.15**	.15**

*Significant at .05 level

**Significant at .01 level

Discussion

On each of the four accuracy measures, the "best judge" and both group judgments are significantly superior to the average of the individuals composing the group. Thus, the major hypothesis of this experiment is confirmed. There is no consistent pattern of significant differences among the first three procedures mentioned above. As would be expected, on the two scores representing the amount of variability in predictions, the "artificial group" mean tends to be lower than the means of the other three procedures. This tendency is significant, however, only for the Interpersonal Accuracy variance score. It is somewhat surprising to find that the "artificial group" is superior to the "best judge" on the Adjective Check List. The interpretation of this finding seems to be that if both other judges disagree with the "best judge," they are more likely to be right than is the "best judge." If, on the other hand, only one of the other judges disagrees with the "best judge," he is more likely to be wrong than is the "best judge."

This study clearly implies that satisfactory ratings are least likely to be obtained from a single individual. In exploring further implications of these results for an operational rating set up, several other considerations enter in. The first of these is that typically the "best judge" would be difficult to select on an a priori basis, and (because of selection error), "best judges" selected a priori would probably score lower than the "best judges" used in this study. Since each of the group procedures produces results roughly equivalent to the "best judge" selected on an after the fact basis, an extensive (and expensive) effort to identify best judges and use them as raters would appear to be unnecessary.

The second consideration involved in applying these results is that by far the most time in this experiment was consumed in arriving at consensus judgments through group discussion, a finding which one would certainly expect to generalize to other situations. Since the "artificial group" procedure produced results as good as or better than the results produced by the consensus judgment, and required much less time, it would appear to be most appropriate when accuracy and time are both considered. Thus, the best procedure for using ratings in many applied situations would be to obtain several independent ratings from different raters for each ratee, and then combine these ratings statistically into a single rating. It should be noted, however, that the superiority of the "artificial group" in terms of time required (and therefore expense) might disappear if only a single summary rating were required rather than the many relatively specific judgments required by the experimental procedure used in this study.

A limitation to these conclusions is the fact that each rater in this experiment was basing his ratings on the same or identical information (i.e., seeing the same movies of the interviews). If different raters are basing their ratings on different information, some other procedure involving the sharing of this information might be superior.

In addition to the practical implications outlined above, these results present, in the opinion of the authors, at least two more basic additions to previous psychological research. The first of these is the demonstration through both the Stereotype Accuracy correlation term and the Interpersonal Accuracy correlation term of the Belief-Values Inventory that accuracy is increased through grouping independent of a reduction in variability (see Table 1). Unlike the other results

of this experiment, this would not necessarily be expected on the basis of previous studies comparing group and individual performance. The second major addition is related to the current controversy in the "interpersonal perception" literature over the relative merits of various different types of accuracy scores (Cronbach, 1955). In the current study the total score on the Adjective Check List, the total score on the Belief-Values Inventory, and the Stereotype Accuracy and Interpersonal Accuracy correlation terms all gave consistent results and, more important, results which make sense in terms of previous research comparing group and individual performance. This would lead one to hope that the interpretations of different types of accuracy scores have more in common than previous investigators have thought.

CHAPTER III

Components of Person Perception Scores and the Clinical and Statistical Prediction Controversy

On the basis of experience with the analysis of components of accuracy scores reported in Chapter I, Richards (1960) has recently suggested a reconceptualization of the clinical and statistical prediction controversy. Briefly, Richards suggests that previous studies comparing clinical and statistical prediction have been heavily loaded on Stereotype Accuracy, and that if comparisons were made on a measure of the Interpersonal Accuracy component, the results might well favor clinical predictions.

A study reported in detail elsewhere (Cline and Richards, 1960 B) was conducted to test this proposed reconceptualization. The specific hypotheses tested were that on the Belief-Values Inventory, statistical prediction is superior to clinical on Stereotype Accuracy, but that clinical prediction is superior to statistical on Interpersonal Accuracy. These hypotheses were tested using 56 college student "clinicians," who were tested in the experimental judging situation in which they made predictions about six standard persons presented by means of sound-color movies of an interview. The data support both hypotheses. An incidental finding was that these student clinicians differentiated much more between interviewees than the accuracy of their differentiations justified.

In the researchers' opinion, these results help to clarify some of the confusing issues raised by Meehl (1954) and others, and further demonstrates the power and utility of component accuracy scores when

these scores are both psychologically meaningful and methodologically sophisticated. It is encouraging to note that this study suggests that clinical prediction is well suited to many usual activities of typical clinicians, e.g. rank ordering a group of patients in terms of probable benefit from psychotherapy.

References

1. Boneau, C. A. The effect of violations of assumptions underlying the t test. Psychol. Bull., 1960, 57, 49-64.

2. Box, G. E. P. Non-normality and tests on variances. Biometrika, 1953, 40, 318-335.

3. Bronfenbrenner, U., Harding, J., & Gallwey, Mary. The measurement of skill in social perception. In D. C. McClellend, Ed. Talent and Society. New York, Van Nostrand, 1958.

4. Cline, V. B. Ability to judge personality assessed with a stress interview and sound film technique. J. abnorm. soc. Psychol., 1955, 50, 183-187.

5. Cline, V. B., & Richards, J. M., Jr. Variables related to accuracy in interpersonal perception. Annual Report, November, 1958. Contract No. NR 171-146, Group Psychology Branch Office of Naval Research. Department of Psychology, University of Utah, Salt Lake City, Utah.

6. Cline, V. B., & Richards, J. M., Jr. Variables related to accuracy in interpersonal perception. Annual Report, November, 1959. Contract No. NR 171-146, Group Psychology Branch, Office of Naval Research. Department of Psychology, Univdrsity of Utah, Salt Lake City, Utah.

7. Cline, V. B., & Richards, J. M., Jr. Accuracy of interpersonal perception--a general trait? J. abnorm. soc. Psychol., 1960, 60, 1-7 (A).

8. Cline, V. B., & Richards, J. M., Jr. Components of person perception scores and the clinical and statistical prediction

controversy: An empirical test of a proposed reconceptualization. Technical Report II. Contract No. NR 171-146, Group Psychology, Branch Office of Naval Research, Department of Psychology, University of Utah, Salt Lake City, Utah, 1960 (B).

9. Cline, V. B., & Richards, J. M., Jr. A note on the generality of accuracy of interpersonal perception. J. abnorm. soc. Psychol., in press.

10. Cline, V. B., & Richards, J. M., Jr. A comparison of individuals vs. groups in judging personality. J. appl. Psychol., in press.

11. Cronbach, L. J. Processes affecting scores on "Understanding of other" and "assumed similarity." Psychol. Bull., 1955, 52, 177-193.

12. Cronbach, L. J., & Gleser, Golding C. Assessing similarity between profiles. Psychol. Bull., 1953, 50, 456-473.

13. Li, J. C. R. Introduction to statistical inference. Ann Arbor, Michigan, Edwards Brothers, 1957.

14. Lorge, I., Fox, D., Davitz, J., & Brenner, M. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. Psychol. Bull., 1958, 55, 337-372.

15. Meehl, P. E. Clinical versus statistical prediction. Minneapolis, University of Minnesota Press, 1955.

16. Richards, J. M. A reconceptualization of the clinical and statistical prediction controversy in terms of components of accuracy of interpersonal perception scores. Technical Report I. Contract No. Nr 171-146, Group Psychology Branch, Office of Naval Research, Department of Psychology, University of Utah, Salt Lake City, Utah, 1960.

Project Staff

1. Principal Investigator--Dr. Victor B. Cline, Ph.D., 1953, University of California, Berkeley. Research Scientist with HumRRO Human research Unit No. 2, Ft. Ord, California 1953-1956. Assistant Research Professor, Department of Psychology, 1957--to present. Principal Investigator delinquency project, 1956-1958, Principal Investigator ONR sponsored research in interpersonal perception, 1957-1960.
2. Research Associate--Dr. James M. Richards, Jr., A.B., Davidson College, 1953; M.A., Emory University, 1955; Ph.D., 1960, University of Utah (Psychology). Psychological Research Associate, U.S. Army Leadership Human Research Unit, 1955-1957. Research Assistant and Associate, Department of Psychology, University of Utah 1957-1960.
3. Statistical Clerk-Typist--Mr. Max Bardin, graduate student, Department of Psychology, University of Utah.