A Method of Estimating Assessment Center Rater Proficiency

1977

Dale Ann Gallo

University of Central Florida

Find similar works at: http://stars.library.ucf.edu/rtd

University of Central Florida Libraries http://library.ucf.edu

Part of the Psychology Commons

STARS Citation


This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
A METHOD OF ESTIMATING
ASSESSMENT CENTER RATER
PROFICIENCY

BY

DALE ANN GALLO
B.S., Florida Southern College, 1973

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science: Psychology
in the Graduate Studies Program of the College of Social Sciences
of Florida Technological University

Orlando, Florida
1977
ACKNOWLEDGMENT

The author would like to express thanks to Dr. Cabot Jaffee, Dr. Wayne Burroughs and Dr. Ed Shirkey for assistance and guidance in completion of this research. A special thanks to my parents, who made it possible to engage in the study of Psychology.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iii</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>METHOD</td>
<td>11</td>
</tr>
<tr>
<td>RESULTS</td>
<td>19</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>58</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>60</td>
</tr>
<tr>
<td>REFERENCE LIST</td>
<td>63</td>
</tr>
</tbody>
</table>
INTRODUCTION

The fundamental aim of all testing is the evaluation of individuals and of individuals' performances on continua representing definable psychological traits and functions. The assessment center method of evaluating individual characteristics and potential, specifically managerial ability, involves the use of situational methods (group exercises, In-Baskets) which are designed to simulate the critical aspects of the job. The personal characteristics selected for evaluation reflect varied aspects of what could be referred to as "criterion" performance. Some are directly related to management functions (e.g., organizing, planning, decision making, problem solving). Others refer to interpersonal relationships and influence (e.g., communications skills, personal impressions, sensitivity) (Bray and Grant, 1966).

A list of skills to be evaluated in this study has been adapted from the initial experimental work of W. D. Storey, of the Management Development Institute (Crotonville, N.Y.) of General Electric Co. as used in Effective Management Selection (Jaffee, 1971).

Much attention has been paid to estimating reliability of ratings, especially interrater reliability. High interrater reliability is generally taken to indicate that the raters have
observed and recorded the same values of the ratee's behavior. "However, there may be common biases among raters" (Guilford, 1954). Among those errors in rating are:

Errors of Leniency . . . These involve raters judging those for whom they feel a personal responsibility more leniently than they should. This is presumably a constant tendency regardless of the trait. Some raters are aware of this and may consequently "bend over backwards" and as a result rate individuals lower than they should. Leniency errors apply to a general constant tendency for a rater to rate too high or too low for whatever reason. When rating is too low, the constant error is one of negative leniency. Positive leniency is by far the most common one. Anticipating this, steps can be taken to help counteract it by avoiding extremely derogatory steps (Guilford, 1954).

Errors of Central Tendency . . . H. L. Hollingworth (1952) is credited with naming and discovering this error. He found that raters hesitate "to go out on a limb" in rating extremely high or low, even when such ratings are justly deserved. This results in a piling up of ratings around the mean of the ratings of a group. This error can be minimized by increasing the number of steps in the middle range and eliminating extreme statements. When an odd number of steps is used this provides the rater with a central or neutral point which can allow for errors of central tendency.
Halo Effect . . . A rater usually allows his general impression to influence his ratings on specific traits. The general impression swamps his judgment of details. A favorable impression toward a worker is likely to lead a superior to give him unduly high ratings in all traits even though he may be below average on several. Methods for minimizing this error are:

1. Rating all individuals on one trait at a time,
2. Reversing order of descriptive anchor phrases periodically,
3. Training raters to be analytical, discussing with raters reasons for divergent ratings so they will arrive at a consensus.

The halo effect is not unlike the stimulus error of psychophysics. It involves irrelevant criteria with which judgments are contaminated. This error can never be avoided but experience has shown where it is most likely found. We can therefore suspect its influence and know where to avoid it.

According to Symonds (1954), it is more prevalent:
1. In a trait that is not easily observed,
2. In a trait that is infrequently discussed or singled out,
3. In a trait not clearly defined,
4. In a trait involving reactions with people,
5. In a trait of high moral importance.

A Logical Error in Rating . . . Newcomb (1931) indicates an error in rating whose effect is not unlike halo error. This is due to the fact that judges are likely to give similar ratings
for traits that seem logically related in the minds of the raters. In the halo effect it is the apparent coherence of qualities in the same individual, whereas in logical errors it is the apparent logical coherence of various traits irrespective of individuals. The latter can be avoided in part by calling for judgments of objectively observable actions rather than abstract and hence semantically overlapping traits. The logical error increases the intercorrelation of traits.

The Contrast Error . . . This is a tendency for a rater to rate others in the opposite direction from his point of reference or internal standards. In psychophysics it is a discrepancy between the judged value of a stimulus and its actual value. An individual's rating may be based on the context of other individuals' performances with whom he is rated at the time or on the basis of the rater having previously rated individuals whose performances were significantly different from the present ratee's performance. Contrast errors have been well documented in research (e.g., Holmes and Berkowitz, 1961, Wexley et. al., 1973).

The Proximity Error . . . Discovered by Stockford and Bissell (1949), like the logical error and contrast error, the proximity error injects undue covariance among rated trait variables. The reason for this spurious correlation is the nearness in space and time for the rating of two traits. Adjacent traits on a rating scale tend to correlate higher than remote ones, their degrees of actual similarity being equal. This error may be counteracted
by placing similar traits further apart and more obviously disparate ones closer together. This error is one reason for not placing much faith in intercorrelations of ratings as intercorrelation of traits (Guilford, 1954).

Interrater reliability is frequently employed to indicate effectiveness of ratings. Interrater reliability computed on two raters of situational exercises using a graphic rating scale ranged from a minimum of $r.53$ to a maximum of $r.70$, using a eight item checklist varying from 7 to 14 items, the range of average correlations were from $r.67$ to $r.90$ (Bray, 1954). Greenwood and McNamara (1967) found that the reasonably high agreement between the different sources of ratings suggests that all were reacting to many of the same aspects of individual performance in both exercises. Evidently, the objective aspects of both exercises are sufficiently apparent to observers and participants alike to influence them similarly in their ratings.

In the ATT study (Bray and Grant, 1966), estimates of the reliabilities of individual assessors were not determined. However, the magnitudes of the communalities obtained from factor analyses indicated that the pooled ratings for many of the variables were reasonably reliable. According to the researchers, the predictions made by the staff are quite accurate. Approximately 80% of those who have advanced to middle management were judged by the assessment center staff as having such potential. The predictors were even more accurate for those who have not
advanced beyond the first level. Most of these men, 95% were judged as lacking in advancement potential . . . regarding the predictive validities of the assessment center method, the situational exercises and the paper and pencil tests are predictive in management, whereas none of the personality questionnaires correlate with the criterion consistently. Justification for the high cost of the assessment center approach can be obtained from the finding that the assessment ratings account for more of the variance than do simple paper and pencil tests or for that matter any single method used (Bray and Grant, 1966).

There is overwhelming evidence of the high interrater reliabilities and the predictive validities associated with the simulation techniques used in assessment centers. In view of the many errors which can effect reliabilities of ratings and indirectly, valid prediction, there seems at first to be a contradiction. The answer lies in the fact that all the studies reporting high reliability and predictive power, utilized trained raters. According to Guilford (1954), various raters experience with ratings tend to show that the most effective method for improving ratings in many ways is to train raters carefully. This applies to all errors mentioned previously. The rater who knows about the existence of the different kinds of errors can be on the lookout for them and can take steps to prevent them. Training that includes practice followed by group discussion seems to be the most effective. This was demonstrated in an experiment by Wexley
et. al. (1973). A series of studies was conducted involving four methods designed to reduce the occurrence of contrast effects. These methods included warning subjects about contrast effects, providing written anchors depicting high and low performance, a combination of the two methods and a fourth method whereby training was administered. The training consisted of a two hour workshop utilizing discussions about high, low and average performance, videotapes depicting the levels of performance which the subjects were asked to rate. Subsequently, the ratings were discussed along with the different types of rating errors. Of the method studied training was the only method to significantly reduce contrast effects.

Many organizations such as the Bell System, Standard Oil of Ohio, I.B.M., and General Electric (to name a few) are currently implementing Assessment Center Activities as a part of their regular personnel services. In order to do so in a reliable and valid manner, a staff of qualified assessors must be selected and trained. It is herein suggested that a measuring instrument is necessary consisting of a pre-training test to determine a base rate before training is begun and a parallel form test to be administered as a post training measure to evaluate the effects of training. A significant difference in the scores on the two separate administrations may be taken as the effect of training. If we are aware of the trainee's present strengths and weaknesses, then a program of training may be developed which emphasizes the
areas in which a majority are weak and cursory in areas where a majority excel. This will provide a training by objectives format (G. S. Odiorne, 1970), where a base rate of performance is quantifiable, thus the amount of change as a result of training may be quantifiably and objectively measured. With the availability of a pre and post training test, it is possible to objectively determine the effectiveness of the assessor training program. By determining an objective (quantifiable) means of measuring assessor candidate base performance, perhaps the optimal type and length of training can be ascertained. When the destination in terms of reaching a specified ability level is known, it should be far easier to assess present level of ability and derive the necessary course of training to reach the desired goal.

The assessment center is only as valid as its assessors are competent in rating.

Without accurate evaluation, any measure is invalid, even precision instruments are nonfunctional if they are not read and used properly.

The purpose of this study is to demonstrate that it is possible to develop an instrument and a methodology for estimating assessment center raters' proficiency level in rating the variables typically represented in the assessment center situational exercise, namely the LGD.
This study will investigate first the significant differences in types of rating errors occurring in three levels of assessment center raters (naive, intermediate and experts). It is hypothesized that, rating errors will be minimized as a function of increased training and experience. This hypothesis is based on data obtained by researchers such as Wexley, et. al. (1972 and 1973) in their investigations in which training significantly reduced contrast effects and studies by other researchers which show that high reliability and predictive power are the result of carefully trained raters (Guilford, 1954; Greenwood and McNamara, 1967).

Secondly, this study was designed to estimate the amount of each kind of constant error (leniency, halo, rater-trait interaction and residual error), within the context of six separate rating situations. The constant errors having been identified were eliminated relative to the expert rating situation. The above were accomplished according to a method devised by Guilford (1954).

In reference to the study conducted by Guilford (1954), the present study differs in three ways. First, Guilford (1954) used only one rating situation in which the ratees were seven scientists in a research organization who were rated on eight traits having to do with creative performance. The three raters were senior researchers who knew their fellows best. The present study involves the rating of two ratees on two separate LGDs, in which
the raters actually observe the behaviors which they rate. It has been demonstrated that trained assessment center raters can only rate accurately a maximum of two ratees (Jaffee, 1971). Therefore, the number of ratees has been reduced. Secondly, the present study differs in that two ratees are rated on five traits by three groups of raters; experts, intermediates and naive. Guilford's (1954) study is concerned with the rating of seven ratees on five traits by three raters on one occasion. Thirdly, the present study attempts to clearly demonstrate the differences between naive raters, intermediate raters and expert raters in the types of rating errors made, and to isolate the specific type of errors which are characteristic of each level group.

In summary, the experiment was designed to assess raters' present ability level at accurately rating the five dimensions correlated with skill in management.

The adjusted ratings of the experts are used as "absolute ratings" or criteria by which to compare subsequent ratings made by the intermediate and naive groups.
METHOD

Subjects

Naive subjects were 23 male and female undergraduate students at Florida Technological University. They were recruited from an Applied Psychology class in which the Assessment Center method had been discussed. Their participation was in partial fulfillment of course requirements. The age range of subjects was from 17 to 50 years, of age.

Intermediate subjects were 23 males who had just completed a three day training course in assessment center rating techniques. Their participation was in partial fulfillment of course requirements. The course consisted of discussion on rating techniques, followed by practice in rating videotaped assessment center exercises. These ratings were subsequently critiqued and discussed.

The expert group was comprised of three Ph.D. level Industrial Psychologists, all of which were consultants on Assessment Center applications in industry.

Instruments

Two LGDs were videotaped. The two LGDs each contained the same two target individuals along with four additional participants used to generate discussion. In the first LCD, one target individual (ratee 1) exhibits less than satisfactory performance while the second target individual (ratee 2) exhibits more than
satisfactory performance. In the second LGD, ratee 1, exhibits more than satisfactory performance, while ratee 2 exhibits only satisfactory performance. Ratee 1 was instructed by the experimenter to perform poorly in the first LGD by interrupting others while speaking, and behaving rudely and sarcastically. Ratee 2 behaved naturally in both LGDs as did ratee 1 in the second LGD.

Both LGDs consisted of three males and three females. All participants were graduate students in the Industrial Psychology program at the University. Each of the LGDs were thirty minutes in duration.

The task in the first LGD was to select an appropriate bicentennial theme for a manufacturing company's lobby. Ten potential topics were presented. The participants were to select the top three in order of suitability. The task in the second LGD was more complex than the first. It involved organizing an aluminum can pickup for charity. Maps depicting the location and amounts of pickups as well as mileage were used along with specifics regarding numbers of trucks, drivers, helpers and time limitations.

Participants were given sufficient time to review instructions prior to initiating discussion. Five minutes for the first LGD and fifteen for the second.

All subjects in the three rating groups (naive, intermediate, expert) were given the same rating form, instructions were attached. Each contained five traits on which the target individuals
were to be rated. These traits were: organization and planning, decision making, leadership, perception, and oral communications. The traits were adopted from a Goals Checklist provided by Jaffee (1971) for the purpose of evaluating LGDs. A separate seven-point graphic rating scale containing the five traits provided for each of the target individuals on each of the two LGDs. A seven-point graphic rating scale was used. See Appendix A for the actual rating forms and instructions used along with definitions of each trait.

Procedure

The task for each of the three groups of subjects (expert, intermediate and naive) was to view two LGDs and to rate the quality of performance exhibited by two target individuals. Videotaped LGDs were used, to approximate actual assessment center situations.

The same seven-point graphic rating scales were used for each of the two LGDs by all three rating groups. The instructions and trait definitions were presented to the subjects in advance ten minutes prior to viewing the videotape. The naive subjects were informed that the skills to be observed and rated were actually represented on the videotape. It was explained that the numbers 1-7 were to designate the degree to which the performers on the tape displayed the behaviors in question. Seven being the highest and one the lowest level. Subjects were then instructed to view the videotape observing the two target individuals as closely
as possible. At the end of each thirty minute videotaped LGD, the subjects were given ten minutes to rate the two target individuals.

The same videotaped LGDs were also viewed by the three expert judges, utilizing the same seven-point graphic rating scale to rate the two targets. The expert ratings were intended as the standard by which the naive and intermediate raters could be compared. The amount of each kind of constant error was estimated and eliminated to derive adjusted ratings free from bias for the expert ratings.

Statistical Analysis

1) The data were treated as two-way factorial designs, fixed effects analyses of variance.

2) Variations were over rater, ratee and trait, the three independent variables.

There was no way of testing for significance of the triple interaction, of rater with trait with ratee, because there were no replications possible within such combinations and hence no way of estimating residual variance (Guilford, 1954).

The present study involved three groups of raters; naive, intermediate and expert with each group rating two target ratees on two separate LGDs. The data is presented in eighteen, two-way factorial designs with replications. Each of the three rating groups (expert, intermediate and naive) require six matrices with three for each of the two LGDs. The three matrices presented
below are used for each of the two LGDs. For the expert rater group, the first matrix, treats ratings as replications across individual ratees and results in the $3 \times 5$ fixed effects ANOVA matrix with $n = 2$ per cell as shown below and used for LGD I and II.

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
 & A & B & C & D & E \\
\hline
1 & & & & & \\
\hline
2 & & & & & \\
\hline
3 & & & & & \\
\hline
\end{array}
\]

Expert Raters

Matrix 1

The second matrix, for the expert rating group, treats ratings as replications across traits and results in the $3 \times 2$ fixed effects ANOVA with $n = 5$ per cell as shown below and used for LGD I and II.

\[
\begin{array}{|c|c|c|}
\hline
 & 1 & 2 & 3 \\
\hline
1 & & & \\
\hline
2 & & & \\
\hline
\end{array}
\]

Expert Raters

Ratees

Matrix 2

The third matrix, for the expert rating group, treats ratings as replications across raters and results in a $2 \times 5$ fixed effects
ANOVA with \( n = 3 \) per cell as shown below and used for LGD I and II.

![Matrix 3 Diagram]

For the intermediate rating group, the fourth matrix, treats ratings as replications across individual ratees and results in the \( 23 \times 5 \) fixed effects ANOVA matrix with \( n = 2 \) per cell, and is used for LGD I and II.

For the intermediate rating group, the fifth matrix treats ratings as replications across traits and results in a \( 2 \times 23 \) fixed effects ANOVA with \( n = 5 \) per cell, and is used for exercises I and II.

The sixth matrix for the intermediate group treats ratings as replications across raters and results in the \( 2 \times 5 \) fixed effects ANOVA matrix with \( n = 23 \) per cell, and is used for both LGD I and II.

The seventh, eighth and ninth matrices for the naive rating groups will be the same as the fourth, fifth and sixth matrices for the intermediate rating groups since there are 23 raters in
both groups. The aforementioned analyses were performed using electronic data processing (USF STATLIB).

According to Guilford (1954), in order to determine the variations of ratings when rater and ratee are combined, a matrix of arithmetic means for rater-ratee combinations was necessary. Each mean in the body of Table 10 was based on five observations of traits. The variations among these means include the simple influences or rater differences and ratee differences which must be removed to show the rater-ratee interaction effects which were found to be significant in LGD II but not in LGD I for the expert raters. To demonstrate this the amounts of each kind of constant error have been estimated for both LGDs. The last column in Table 10 shows deviations of rater means from the grand mean. The deviations (from the grand mean) were the rater's constant errors. The last two rows of Table 10 show individual differences over all traits and raters. In LGD I, the deviations $d_i$ (from the grand mean, 4.4) of individual means, in all traits combined may include halo effects which all raters agree upon. The interaction errors that have been estimated are actually relative halo effects. That is the subjective contributions to the ratings within the context of these raters, traits and ratees.

In Table 11 are the adjusted means of the rater-ratee combinations. The adjustment eliminates the inter-rater differences and inter-ratee differences.
From the mean in each cell of Table 10 the corresponding deviations \( X'_{dl} \) and \( d_i \) were deducted. This ensures that the adjusted means for all raters and for all ratees will equal the grand mean.

In order to estimate interaction errors of the rater-trait type, the process was similar to the above. The steps were analogous. Individual differences among raters were ignored, combinations of raters and traits were averaged.

The deviations of the adjusted means from the grand mean are given in Table 12. These deviations are relative halo effects.

The Pearson product-moment correlation \((r)\) was used to determine if there was a relationship between the experts' ratings on traits. In other words, to determine the correlations between the five traits as demonstrated by the experts' ratings. Two tailed t-tests were used to determine if correlations between each trait were significant at the .001 level. The above methods were performed by electronic data processing using SPSS.

Pearson product-moment correlation coefficients were also computed to determine if there were intercorrelations between the ratings of traits on LGD I and the ratings of traits on LGD II. These coefficients were computed separately for each of the three levels of raters and for each individual ratee.
RESULTS

All Tables and Figures referred to in the text are listed in the back of this section. In Tables 1 - 9, are the summaries of eighteen analysis of variance solutions. That is six solutions for each of the three rating groups (expert, intermediate and naive). The data are treated as two-way factorial designs treating ratees, raters and traits as replications.

Figure 1 graphically depicts the ratee means given by the expert group on LGD I. It can be seen that rater 2 has overrated ratee 2 and underrated ratee 1, in relation to the other two raters. In Figure 2, are the ratee means by the experts on LGD II. It can be seen that again rater 2 has overrated ratee 2 and underrated ratee 1 in relation to the other two raters. The intersection of the lines demonstrates an interaction effect between rater and ratee. This interaction effect was found to be significant for LGD II in the ANOVA solution which will be discussed in detail further ahead.

Figure 3 represents the ratee means given by intermediate raters on LGD I. The general tendency shown indicates that ratee 2 tended to receive higher ratings than ratee 1, however, the intersections demonstrate an interaction between rater and ratee for several raters. This interaction was found to be significant in the subsequent ANOVA. Figure 4 represents the ratee means
given by intermediate raters on LGD II. It can be seen that the ratings of the two ratees were much closer with rater-ratee interaction, which again proved to be significant on a subsequent ANOVA.

Figure 5 and 6 show the ratee means given by naives on LGD I and II. On LGD I, ratee 1 received far lower ratings than ratee 2, no interaction occurred. On LGD II, Figure 6, the two ratees were rated almost equally, with a significant interaction shown graphically in the Figure and in the subsequent ANOVA.

Refer to Table 1 - 3 for the expert raters, 4 - 6 for the intermediate raters and 7 - 9 for the naive raters for the two-way fixed effects ANOVAs. No significant difference was found between traits for any rating group on LGD I or II. No significant interaction was found for any group between rater and trait nor between ratee and trait.

The F ratios for the experts indicate a significant difference between ratees on both LGD I and II, when treating raters and traits as replications. The F ratios for the intermediate group also indicates a significant difference between ratees in LGD I and II, when treating raters and traits as replications.

The F ratios for the naive group indicate significant differences between ratees only on LGD I, but no significant difference between ratees on LGD II.

A possible explanation could be that the experts and the trained intermediate raters were able to discriminate between the performances of the ratees even when their performances were
similar on LGD II. The naive group, however, was able to discern between the ratees' performances only when the disparities were the most obvious on LGD I. They could not make the finer discriminations which were restricted to the trained groups.

The F ratios for the expert group indicate no significant differences between raters in either solution on LGD I or II. The F ratios for the intermediates correspondingly indicate no significant differences between raters in LGD I or II. The differences between raters for the naive group were significant in all solutions for LGD I and II, except a nonsignificant F was found when treating ratees as replications on LGD I.

This would seem to indicate that the experts and trained intermediates rated more consistently as a group, whereas the naive raters had little agreement and widely disparate ratings for LGD II. The nonsignificant F in LGD I (also shown graphically in Figure 5) is probably due to the apparent large discrepancies between the ratees on LGD I. This made it possible for even the naive group to clearly distinguish between the two ratees. Therefore, the naive raters were in agreement and showed no significant difference between raters for LGD I. Differences between raters indicate variances from rater to rater in relative leniency errors (Guilford, 1954). According to the ANOVAs the expert and intermediate raters were relatively free from leniency errors, that is, the tendency to over or under value ratees in general.
Of the simple interaction variances that between rater and trait was not significant for experts, intermediates or naives on LGD I or II, according to the ANOVAs. According to Guilford (1954), the rater-trait interaction represents a general tendency to over-value or under value a certain trait in others. This could be interpreted as a logical error in rating, if it were significant.

The interaction variance between rater and ratee was not significant in LGD I for the experts in LGD I, however, it was significant in LGD II (F = 3.5 p < .05, 2 df). This may be seen in Figure 2. Rater 2 tends to under value ratee 1 and over value ratee 2, in relation to the ratings of the other two raters. The rater-ratee interaction was significant on both LGD I and II for the intermediate raters.

The rater-ratee interaction was not significant on LGD I for the naive raters but was significant on LGD II.

Concerning the intermediate raters, as a group, apparently the gap in performance between the two ratees was not perceived to be as great by the intermediates as it was by the naives on the first LGD. On LGD II when the ratees' performances became more similar the ratings of the intermediate and naive raters resulted in the interaction effects shown in Figures 4 and 6.

According to Guilford (1954), the total error can be broken down into components. Some of these components are identified as follows:
\( X_{kl} = \) rater K's "leniency error"
\( X_{ki} = \) rater K's "halo error"
\( X_{kj} = \) raters' rater-trait interaction error
\( X_{ijkr} = \) a residual error

Total error is a summation of the above components.

\( X_{ijk} = \) a rating of person I in trait j by rater k
\( X_{ijt} = \) the true value of person I in trait j
\( X_{ijk} = X_{ijt} + X_{kl} + X_{kj} + X_{ijkr} \)

Any rating is thus composed of the above true and error variance.

The amounts of each kind of constant error in the preceding equation have been estimated for each individual expert rater.

To see the results of this process refer to Tables 10 - 15. Each mean in Table 10 is based on two observations, one for each ratee. The variations include influences of rater and ratee differences. These differences were removed in Table 11 to find the contribution of interaction effects. In Table 10, the last column \( X'_{kl} \) depicts the amount of the raters' constant error. Rater 1 had no constant error \( X'_{kl} = .00 \) across ratees. Rater 2 had \( X'_{kl} = -.20 \) and rater 3 had \( X'_{kl} = +.20 \) across all ratees. The last two rows of Table 10, indicate difference from the grand mean (4.40) across all traits and raters. These deviations (\( di = -1.20 \) for ratee 1 and \( di = +1.20 \) for ratee 2), may include halo effects on which all raters agree. Referring to Table 11, it can be seen that four of the ratings still deviate from the grand mean even
after adjustment for leniency and rater deviations. These deviations although quite small, could be the result of interaction effects between rater and ratee, that is, halo effect. In LGD I, these interactions between rater and ratee were found to have a nonsignificant F ratio. In Table 12 it is seen that these errors are small and range from $X'^{ki} = +.20$ to $-.20$ with all deviations summing to zero. Therefore, the nonsignificant F ratio for rater-ratee interaction was reflected by the halo error adjustment process.

Tables 13 - 15 display the results of the adjustment for rater-trait interaction. Means were derived from ratings by rater combined with trait in Table 13. It can be seen that rater deviations (dt) are the greatest in trait A = $-.40$, trait C = $-.80$, trait D = $+.43$ and trait E = $+.60$. These deviations may be the result of actual differences in the traits being rated. In Table 14 are the corrected ratings with the deviations removed. The amount of deviation from the grand mean is quite small. In Table 15 the actual amounts of interaction between rater and trait seem to zero across all traits for each rater 1, 2 and 3. Traits D and C, however still contain $+.20$ and $-.20$ amounts of contrast error. This rater-trait interaction was found to be nonsignificant in the ANOVA for experts in LGD I and can be disregarded.

Tables 16 - 21 give the same types of information as Tables 10 - 15, however, the former tables refer to the expert rating group in relation to LGD II. The same adjustment process applies to
Tables 16 and 17. Advancing to Table 18, the amount of rater and ratee interaction is estimated, \((X'_{kl})\) or halo). It can be seen that Rater 2 undervalued ratee 1, \(X'_{kl} = -0.40\) and overvalued ratee 2, \(X'_{kl} = +.40\). Rater 3, overvalued ratee 1, \(X'_{kl} = +.30\) and undervalued ratee 2, \(X'_{kl} = -.30\). There was found a significant interaction between ratee and rater in the ANOVA solution for LGD II.

Advancing to Table 21, when raters are combined with traits to measure the contribution of rater and trait interaction, the greatest amount of deviation from the mean occurs. On LGD II, Trait C and E both contribute, \(X'_{kj} = +.30\). Since the rater-trait interaction was nonsignificant, this may be disregarded.

In order to determine the intercorrelation of the traits, Pearson product-moment correlation coefficients were computed for each rater group (expert, intermediate and naive). Each possible trait combination was intercorrelated in Table 22 and 23, for LGD I and II respectively. On LGD I all possible trait combinations were significantly correlated for the naive group. On LGD II only one trait combination was nonsignificant probably due to random chance.

On LGD I, the intermediate groups ratings of the traits resulted again in the significant intercorrelation of all traits. Possibly, even the trained intermediates are influenced by logical errors, in rating traits which they perceive to be logically related. The intermediate group had significant intercorrelations for all trait combinations except for three trait combinations on
LGD II. On LGD I, the experts' had six nonsignificant trait combination intercorrelations and two nonsignificant intercorrelations on LGD II.

Regarding the naive and intermediate groups with \( n = 23 \), it is possible that either the traits are in fact logically related and should be intercorrelated or the raters are systematically affected by halo which tends to increase intercorrelation of traits.

Regarding the experts, with \( n = 3 \) it is not appropriate to draw conclusions since the power of any test is a function of sample size.

Returning back to the formula for adjusting raw score ratings, to illustrate the process, the rating of LGD I, individual ratee 1 in trait A by expert rater 1 follows:

\[
X_{ijk} - X'_{kl} - X'_{ki} - X'_{kj} = X'_{ijk}
\]

\[
X_{1Al} = 2.00 - .00 - .00 - .00 = 2.00
\]

In essence this rating was actually free from constant error. Another example taken from LGD II for individual ratee 1 in trait A by expert rater 2:

\[
X_{ijk} - X'_{kl} - X'_{ki} - X'_{kj} = X'_{ijk}
\]

6.00 - (+.27) - (-.40) - (+.60) = 5.53

Hence, rater 2 overvalued ratee 1 in trait A by +.47 on the second LGD. We know that rater-ratee interaction was significant from the ANOVA.
Pearson product-moment correlation coefficients were computed to determine the relationship between a rating groups' rating of an individual ratee on the first LGD in relation to the same groups' ratings on the second LGD. The results in Table 24, indicate no significant correlations between ratings on the first LGD or the second LGD for the intermediate group when rating ratee 1 or ratee 2.

In the naive rating group three out of five traits on each LGD were found to be significantly correlated with subsequent ratings of the same trait.

Both naive and intermediate raters had the same sample size $n = 23$.

Intercorrelations of traits between LGD I and II were also computed for the expert group. No significant correlations were found, however, due to the small sample size, $n = 3$, no conclusions will be drawn.
Figure 1
Profiles of Rating Means for Experts LGD I
Figure 2

Profiles of Rating Means for Experts LGD II

---

Ratee 1

Ratee 2

Expert Raters
Figure 3

Profiles of Ratings Means for Intermediates LGD I

Ratee 1

Ratee 2
Figure 4
Profile of Rating Means for Intermediates LGD II

![Profile of Rating Means for Intermediates LGD II](image)
Figure 5
Profiles of Rating Means for Naives LCD I
Figure 6
Profiles of Rating Means for Naives LGD II

Ratee 1
Ratee 2
Table 1

Summary of Analysis of Variance of Ratings of Two Ratees in Five Traits As Given By Three Expert Raters On LGD I and II

Treating Rater Differences As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>df</th>
<th>MS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratees(I)</td>
<td>43.2</td>
<td>4.8</td>
<td>1</td>
<td>1</td>
<td>43.2</td>
<td>4.8</td>
</tr>
<tr>
<td>Traits(T)</td>
<td>7.53</td>
<td>.8</td>
<td>4</td>
<td>4</td>
<td>1.88</td>
<td>.2</td>
</tr>
<tr>
<td>I x T</td>
<td>.466</td>
<td>2.4</td>
<td>4</td>
<td>4</td>
<td>.12</td>
<td>.6</td>
</tr>
<tr>
<td>Within</td>
<td>20.00</td>
<td>9.46</td>
<td>20</td>
<td>20</td>
<td>1.00</td>
<td>.47</td>
</tr>
<tr>
<td>Total</td>
<td>71.20</td>
<td>17.46</td>
<td>29</td>
<td>29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05
Table 2

Summary of Analysis of Variance of Ratings of Two Ratees in Five Traits As Given By Three Expert Raters On LGD I and II

Treating Differences Between Traits As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Ratees(I)</td>
<td>43.2</td>
<td>4.8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Raters(R)</td>
<td>.2</td>
<td>1.3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>R x I</td>
<td>1.4</td>
<td>2.6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Within</td>
<td>26.4</td>
<td>8.8</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>71.2</td>
<td>17.5</td>
<td>29</td>
<td>29</td>
</tr>
</tbody>
</table>

*p < .05
Table 3

Summary of Analysis of Variance of Ratings of Two Ratees in Five Traits As Given By Three Expert Raters On LGD I and II

Treating Differences Between Ratees As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Raters(R)</td>
<td>.2</td>
<td>1.3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Trait(T)</td>
<td>7.53</td>
<td>.8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R x T</td>
<td>13.13</td>
<td>3.87</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Within</td>
<td>50.34</td>
<td>11.5</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

*p< .05
Table 4

Summary of Analysis of Variance of Ratings of Two Raters in Five Traits As Given By Twenty-Three Intermediate Raters On LGD I and II

Treating Rater Differences As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Ratees(I)</td>
<td>104.5</td>
<td>21.92</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Traits(T)</td>
<td>5.98</td>
<td>2.48</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R x T</td>
<td>9.57</td>
<td>6.84</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Within</td>
<td>230.23</td>
<td>232.50</td>
<td>221</td>
<td>221</td>
</tr>
<tr>
<td>Total</td>
<td>350.28</td>
<td>263.74</td>
<td>230</td>
<td>230</td>
</tr>
</tbody>
</table>

*p < .05
Table 5

Summary of Analysis of Variance of Ratings of Two Raters in Five Traits As Given By Twenty-Three Intermediate Raters On LGD I and II

Treating Differences Between Traits As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Ratees(I)</td>
<td>104.5</td>
<td>1</td>
<td>104.5</td>
<td>209*</td>
</tr>
<tr>
<td>Raters(R)</td>
<td>33.76</td>
<td>22</td>
<td>1.53</td>
<td>3.06</td>
</tr>
<tr>
<td>(R x I)</td>
<td>120.4</td>
<td>22</td>
<td>5.47</td>
<td>10.94*</td>
</tr>
<tr>
<td>Within</td>
<td>91.62</td>
<td>185</td>
<td>.50</td>
<td>.61</td>
</tr>
<tr>
<td>Total</td>
<td>350.28</td>
<td>230</td>
<td>.50</td>
<td>.61</td>
</tr>
</tbody>
</table>

*p < .05
Table 6

Summary of Analysis of Variance of Ratings of Two Raters in Five Traits As Given By Twenty-Three Intermediate Raters On LGD I and II

Treatng Differences Between Ratees As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Raters(R)</td>
<td>33.76</td>
<td>51.46</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Traits(T)</td>
<td>5.98</td>
<td>2.48</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>(I x T)</td>
<td>39.02</td>
<td>49.32</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>Within</td>
<td>271.52</td>
<td>160.48</td>
<td>116</td>
<td>116</td>
</tr>
<tr>
<td>Total</td>
<td>350.28</td>
<td>263.74</td>
<td>230</td>
<td>230</td>
</tr>
</tbody>
</table>

*p < .05
Table 7

Summary of Analysis of Variance of Ratings of Two Ratees in Five Traits As Given By Twenty-Three Naive Raters On LGD I and II

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Ratees(I)</td>
<td>353.2</td>
<td>.435</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Traits(T)</td>
<td>10.65</td>
<td>6.1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R x T</td>
<td>9.87</td>
<td>6.95</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Within</td>
<td>295.2</td>
<td>227</td>
<td>220</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td>668.90</td>
<td>240.8</td>
<td>229</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>353.2</td>
<td>.435</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2.66</td>
<td>1.53</td>
<td>4.99</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>2.47</td>
<td>1.74</td>
<td>1.73</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>1.34</td>
<td>1.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*<p < .05
Table 8

Summary of Analysis of Variance of Ratings of Two Ratees in Five Traits As Given By Twenty-Three Naive Raters On LGD I and II

Treating Differences Between Traits As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
</tr>
<tr>
<td>Ratees(I)</td>
<td>353.2</td>
<td>.435</td>
<td>1</td>
</tr>
<tr>
<td>Raters(R)</td>
<td>108.0</td>
<td>89.4</td>
<td>22</td>
</tr>
<tr>
<td>(R x I)</td>
<td>59.3</td>
<td>49.4</td>
<td>22</td>
</tr>
<tr>
<td>Within</td>
<td>148.4</td>
<td>101.6</td>
<td>184</td>
</tr>
<tr>
<td>Total</td>
<td>668.9</td>
<td>240.8</td>
<td>229</td>
</tr>
</tbody>
</table>

*p < .05
Table 9

Summary of Analysis of Variance of Ratings of Two Ratees in Five Traits As Given By Twenty-Three Naive Raters On LGD I and II

Treating Differences Between Ratees As Replications

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Raters(R)</td>
<td>108.0</td>
<td>89.4</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Traits (T)</td>
<td>10.65</td>
<td>6.1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>(I x T)</td>
<td>63.74</td>
<td>42.3</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>Within</td>
<td>486.5</td>
<td>103</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>Total</td>
<td>668.9</td>
<td>240.8</td>
<td>229</td>
<td>229</td>
</tr>
</tbody>
</table>

*p < .05
Table 10

Estimations of the Contributions to the Interactions of Rater With Ratee Among the Ratings of Two Ratees by Three Raters, On Five Traits for LGD I

Means of Raters Derived From Ratings By Different Raters

<table>
<thead>
<tr>
<th>Ratees</th>
<th>1</th>
<th>2</th>
<th>All Ratees</th>
<th>$X'_{k1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.40</td>
<td>5.40</td>
<td>4.40</td>
<td>-.00</td>
</tr>
<tr>
<td>2</td>
<td>2.80</td>
<td>5.60</td>
<td>4.20</td>
<td>-.20</td>
</tr>
<tr>
<td>3</td>
<td>3.40</td>
<td>5.80</td>
<td>4.60</td>
<td>+.20</td>
</tr>
<tr>
<td>All Raters</td>
<td>3.20</td>
<td>5.60</td>
<td>4.40</td>
<td>.00</td>
</tr>
<tr>
<td>d1</td>
<td>-1.20</td>
<td>+1.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 11

Estimations of the Contributions to the Interactions of Rater With Ratee Among the Ratings of Two Ratees by Three Raters, On Five Traits for LGD I

Means Corrected for Rater Error $X'_{kl}$, and for Ratee Deviations $d_i$

<table>
<thead>
<tr>
<th>Ratees</th>
<th>1</th>
<th>2</th>
<th>All Ratees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.60</td>
<td>4.20</td>
<td>4.40</td>
</tr>
<tr>
<td>2</td>
<td>4.20</td>
<td>4.60</td>
<td>4.40</td>
</tr>
<tr>
<td>3</td>
<td>4.40</td>
<td>4.40</td>
<td>4.40</td>
</tr>
<tr>
<td>All Raters</td>
<td>4.40</td>
<td>4.40</td>
<td></td>
</tr>
</tbody>
</table>
Table 12

Estimations of the Contributions to the Interactions of Rater With Ratee Among the Ratings of Two Ratees by Three Raters, On Five Traits for LGD I

Contributions of Interactions of Rater and Ratee:
Halo Errors $X'_{ki}$

<table>
<thead>
<tr>
<th>Ratees</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>+.20</td>
<td>-.20</td>
</tr>
<tr>
<td>2</td>
<td>-.20</td>
<td>+.20</td>
</tr>
<tr>
<td>3</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

\[ \text{.00} \]
Table 13

Estimation of the Contributions to the Interactions of Rater With Trait Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD I

Means of Ratings by Rater Combined With Trait

<table>
<thead>
<tr>
<th>Traits</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>All Rates</th>
<th>X'kl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.0</td>
<td>4.5</td>
<td>3.5</td>
<td>5.0</td>
<td>5.0</td>
<td>4.4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4.0</td>
<td>4.5</td>
<td>3.5</td>
<td>4.5</td>
<td>5.0</td>
<td>4.3</td>
<td>-.1</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>4.5</td>
<td>4.0</td>
<td>5.0</td>
<td>5.0</td>
<td>4.5</td>
<td>+.1</td>
</tr>
<tr>
<td>All Raters</td>
<td>4.0</td>
<td>4.5</td>
<td>3.6</td>
<td>4.83</td>
<td>5</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>dt</td>
<td>-.40</td>
<td>+.10</td>
<td>-.80</td>
<td>+.43</td>
<td>+.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 14

Estimation of the Contributions to the Interactions of Rater With Trait Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD I

Means By Rater and Trait, Corrected for Rater Error $X'_{kl}$ and For Trait Deviations $d_t$

<table>
<thead>
<tr>
<th>Traits</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>All Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.4</td>
<td>4.4</td>
<td>4.3</td>
<td>4.5</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>2</td>
<td>4.5</td>
<td>4.5</td>
<td>4.4</td>
<td>4.1</td>
<td>4.5</td>
<td>4.4</td>
</tr>
<tr>
<td>3</td>
<td>4.3</td>
<td>4.3</td>
<td>4.7</td>
<td>4.4</td>
<td>4.3</td>
<td>4.4</td>
</tr>
<tr>
<td>All Raters</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
<td>4.4</td>
</tr>
</tbody>
</table>
Table 15

Estimation of the Contributions to the Interactions of Rater With Trait Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LCD I

Contributions of Interactions of Rater and Trait;
Contrast Error $X'_{kj}$

<table>
<thead>
<tr>
<th>Traits</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.00</td>
<td>.00</td>
<td>-.10</td>
<td>+.10</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>+.10</td>
<td>+.10</td>
<td>.00</td>
<td>-.30</td>
<td>+.10</td>
</tr>
<tr>
<td>3</td>
<td>-.10</td>
<td>-.10</td>
<td>+.30</td>
<td>.00</td>
<td>-.10</td>
</tr>
<tr>
<td></td>
<td>.00</td>
<td>.00</td>
<td>+.20</td>
<td>-.20</td>
<td>.00</td>
</tr>
</tbody>
</table>
Table 16

Estimation of the Contribution to the Interactions of Rater With Ratee Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD II

Means of Raters Derived From Ratings By Different Raters

<table>
<thead>
<tr>
<th>Ratees</th>
<th>1</th>
<th>2</th>
<th>All Rateses</th>
<th>X'_kl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.60</td>
<td>4.60</td>
<td>5.10</td>
<td>-.03</td>
</tr>
<tr>
<td>2</td>
<td>5.40</td>
<td>5.40</td>
<td>5.40</td>
<td>+.27</td>
</tr>
<tr>
<td>3</td>
<td>5.60</td>
<td>4.20</td>
<td>4.90</td>
<td>-.23</td>
</tr>
<tr>
<td>All Raters</td>
<td>5.53</td>
<td>4.73</td>
<td>5.13</td>
<td></td>
</tr>
<tr>
<td>d1</td>
<td>+.40</td>
<td>-.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 17

Estimation of the Contribution to the Interactions of Rater With Ratee Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD II

Means Corrected for Rater Error $X'_{kl}$, and For Ratee Deviations $d_i$

<table>
<thead>
<tr>
<th>Ratees</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>All Ratees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.23</td>
<td>5.03</td>
<td>5.13</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.73</td>
<td>5.53</td>
<td>5.13</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5.43</td>
<td>4.83</td>
<td>5.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.13</td>
<td>5.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 18

Estimation of the Contribution to the Interactions of Rater with Ratee Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD II

Contributions of Interactions of Rater and Ratee: Halo Errors $X_{kl}$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratees</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>+.10</td>
<td>-.10</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>-.40</td>
<td>+.40</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>+.30</td>
<td>-.30</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note: The table shows the contributions of interactions of rater and ratee for Halo Errors $X_{kl}$.
Table 19

Estimation of the Contributions to the Interactions of Rater With Trait Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD II

Means of Ratings by Rater Combined With Trait

<table>
<thead>
<tr>
<th>Traits</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>All Traits</th>
<th>X'_{kl}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.0</td>
<td>5.0</td>
<td>5.5</td>
<td>5.0</td>
<td>5.0</td>
<td>5.1</td>
<td>- .10</td>
</tr>
<tr>
<td>2</td>
<td>6.0</td>
<td>5.5</td>
<td>5.5</td>
<td>5.5</td>
<td>4.5</td>
<td>5.4</td>
<td>+ .20</td>
</tr>
<tr>
<td>3</td>
<td>4.5</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>4.9</td>
<td>- .30</td>
</tr>
<tr>
<td>All Raters</td>
<td>5.2</td>
<td>5.2</td>
<td>5.3</td>
<td>5.2</td>
<td>4.8</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>di</td>
<td>.00</td>
<td>.00</td>
<td>+.10</td>
<td>.00</td>
<td>-.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 20

Estimation of the Contributions to the Interactions of Rater With Trait Among the Ratings of Two Ratees by Three Expert Raters on Five Traits for LGD II

Means By Rater and Trait, Corrected for Rater Error $X'_{kl}$ and for Trait Deviations

<table>
<thead>
<tr>
<th>Traits</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>All Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.5</td>
<td>5.1</td>
<td>5.5</td>
<td>5.2</td>
</tr>
<tr>
<td>2</td>
<td>5.8</td>
<td>5.3</td>
<td>5.2</td>
<td>5.3</td>
<td>4.7</td>
<td>5.2</td>
</tr>
<tr>
<td>3</td>
<td>4.8</td>
<td>5.3</td>
<td>5.2</td>
<td>5.3</td>
<td>5.7</td>
<td>5.2</td>
</tr>
<tr>
<td>All Raters</td>
<td>5.2</td>
<td>5.2</td>
<td>5.2</td>
<td>5.2</td>
<td>5.2</td>
<td>5.2</td>
</tr>
</tbody>
</table>
Table 21

Estimation of the Contributions to the Interactions of Rater
With Trait Among the Ratings of Two Ratees by Three
Expert Raters on Five Traits for LGD II

Contributions of Interactions of Rater and Trait;
Contrast Error $X_{kj}^r$

<table>
<thead>
<tr>
<th>Traits</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-.10</td>
<td>-.10</td>
<td>+.30</td>
<td>-.10</td>
<td>+.30</td>
</tr>
<tr>
<td>2</td>
<td>+.60</td>
<td>+.10</td>
<td>.00</td>
<td>+.10</td>
<td>-.50</td>
</tr>
<tr>
<td>3</td>
<td>-.40</td>
<td>+.10</td>
<td>.00</td>
<td>+.10</td>
<td>+.50</td>
</tr>
<tr>
<td></td>
<td>+.10</td>
<td>+.10</td>
<td>+.30</td>
<td>+.10</td>
<td>+.30</td>
</tr>
</tbody>
</table>
Table 22

Summary of Intercorrelations Between Traits on LGD I For Experts, Intermediate and Naive Group Raters

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.96</td>
<td>.89</td>
<td>.32*</td>
<td>.33*</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.70</td>
<td>.71</td>
<td>.74</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.81</td>
<td>.75</td>
<td>.71</td>
<td>.62</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.96</td>
<td>.97</td>
<td>.56*</td>
<td>.58*</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.70</td>
<td>.74</td>
<td>.68</td>
<td>.62</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.81</td>
<td>.79</td>
<td>.77</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.89</td>
<td>.97</td>
<td>.69*</td>
<td>.70*</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.71</td>
<td>.74</td>
<td>.64</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.75</td>
<td>.79</td>
<td>.80</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.32*</td>
<td>.56*</td>
<td>.69*</td>
<td>.96</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.74</td>
<td>.68</td>
<td>.64</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.71</td>
<td>.77</td>
<td>.80</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.33*</td>
<td>.58*</td>
<td>.70*</td>
<td>.96</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.77</td>
<td>.62</td>
<td>.67</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.62</td>
<td>.72</td>
<td>.75</td>
<td>.72</td>
<td></td>
</tr>
</tbody>
</table>

*Not Significant at .001

T1 = Organization & Planning
T2 = Decision Making
T3 = Perception
T4 = Leadership
T5 = Oral Communication
### Table 23

**Summary of Intercorrelations Between Traits On LGD II For Experts, Intermediate and Naive Group Raters**

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.49</td>
<td></td>
<td></td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.73</td>
<td>.54</td>
<td></td>
<td>.10*</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.49</td>
<td>.74</td>
<td>.50</td>
<td>.18*</td>
<td></td>
</tr>
<tr>
<td><strong>T2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.49</td>
<td></td>
<td>.45</td>
<td>.39*</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.73</td>
<td></td>
<td></td>
<td>.48</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.49</td>
<td>.66</td>
<td></td>
<td>.39*</td>
<td></td>
</tr>
<tr>
<td><strong>T3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.54</td>
<td>.91</td>
<td></td>
<td>.66</td>
<td>.00*</td>
</tr>
<tr>
<td>I</td>
<td>.74</td>
<td>.66</td>
<td></td>
<td>.48</td>
<td>.39*</td>
</tr>
<tr>
<td>N</td>
<td>.28*</td>
<td>.44</td>
<td>.55</td>
<td>.51</td>
<td>.61</td>
</tr>
<tr>
<td><strong>T4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.49</td>
<td>.79</td>
<td>.66</td>
<td></td>
<td>.32</td>
</tr>
<tr>
<td>I</td>
<td>.50</td>
<td>.51</td>
<td>.48</td>
<td></td>
<td>.44</td>
</tr>
<tr>
<td>N</td>
<td>.45</td>
<td></td>
<td></td>
<td></td>
<td>.61</td>
</tr>
<tr>
<td><strong>T5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>.42</td>
<td>.00*</td>
<td>.00*</td>
<td>.44</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.10*</td>
<td>.18*</td>
<td>.39*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>.39*</td>
<td>.53</td>
<td>.61</td>
<td>.61</td>
<td></td>
</tr>
</tbody>
</table>

*Not Significant at .001

T₁ = Organization and Planning  
T₂ = Decision Making  
T₃ = Perception  
T₄ = Leadership  
T₅ = Oral Communication
Table 24

Summary of the Intercorrelation of Ratings on LGD I and Ratings on LGD II

<table>
<thead>
<tr>
<th>Traits</th>
<th>Ratee 1</th>
<th>Ratee 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive $r_{xy}$</td>
<td>Intermediate $r_{xy}$</td>
</tr>
<tr>
<td>1</td>
<td>.17</td>
<td>-.12</td>
</tr>
<tr>
<td>2</td>
<td>.40*</td>
<td>-.40</td>
</tr>
<tr>
<td>3</td>
<td>.53</td>
<td>.01</td>
</tr>
<tr>
<td>4</td>
<td>.17</td>
<td>-.04</td>
</tr>
<tr>
<td>5</td>
<td>.42*</td>
<td>.20</td>
</tr>
</tbody>
</table>

*p < .05

$x =$ Ratings on LGD I

$y =$ Ratings on LGD II
DISCUSSION

Two findings are of considerable interest. First, trained intermediate raters were able to discriminate between ratees on performance even when levels of performance were quite similar. The ANOVAs for experts and intermediate raters indicate the differences between ratees were significant on both LGD I and II.

Naive raters were unable to discriminate between ratees when level of performance was similar. The difference between ratees was not significant for naive raters ($F = .435 p < .05, 1 df$) on LGD II. There was no significant difference between raters on LGD I or II for intermediates or experts. There was a significant difference between raters on LGD I and II for the naive raters.

Differences between raters indicates variations from rater to rater in leniency errors (Guilford, 1954). Therefore, variations from rater to rater in leniency error were present in the naive group but not in the expert or intermediate group. These findings correspond to data obtained by other researchers in that training can minimize rating errors (Wexley, et al., 1972; Guilford, 1954; Greenwood and McNamara, 1967).

The correlation coefficients computed to determine the relationship between ratings on the first LGD in relation to ratings on the second LGD seem to indicate that trained intermediates did not have the tendency to carry over previous ratings to a new
rating situation. Possibly training served to reduce the number of correlated ratings on traits from one exercise to the next. Perhaps a type of response set or contrast effect is more present in naive raters.

Second, it has been shown that it is possible to detect the presence of leniency, halo and contrast error through the significant interactions of rater-ratee, rater-trait and ratee-trait in two-way factorial ANOVAs. Once significance has been determined relative to a particular rating situation, the amount of interaction error present can be eliminated from an individual raw score. Although the adjustment process was demonstrated only for the experts, the presence of significant interaction in the other groups serves as guide in applying the adjustment. Regarding the Assessment Center method, the adjusted expert ratings could be utilized as absolute ratings by which to compare the ratings of other assessors. Thereby, a videotaped situational exercise such as this could be used as a pre and post training test for Assessment Center raters.

The Assessment Center is designed to minimize rater error. Frequently, an individual ratee is rated by a different rater on each exercise to avoid the carry over of response sets and contrast effect. In the present study, an effort was made to maximize the occurrence of rater error for its subsequent measurement. This was done by using the same ratees in both exercises, thereby increasing the probability of systematic bias.
APPENDIX A
Assessment Center Rater Proficiency Test

Name: ________________________

Date: ________________________

Using a rating scale ranging from weak to outstanding for a particular skill, consider candidates two and five in regards to their performance in the two group exercises you are about to observe. After arriving upon your ratings for each skill, defend your ratings based on incidents you have observed and noted.

Definitions of the skills to be rated:

Organization and planning -- The ability to be well organized and systematic. Plans before taking action. Breaks a task into elements that make it easier to handle. Makes organizing or planning suggestions for the group.

Decision making ---------- The ability to make effective and reasonable decisions concerning people and task while remaining within task boundaries. Gathers available material and makes decisions when they are called for.

Perception --------------- The ability to perceive the critical elements of problems involving people as well as problems involving the task itself. Notes the critical dimensions of the task.

Leadership --------------- The ability to elicit cooperation while remaining task oriented moving the group effectively towards completion of the task.

Oral Communications ......... The ability to communicate ideas effectively and with verbal facility be understood.
Rating Scale Period I Person 2. Circle the appropriate number which represents your opinion of the individual on each skill.

**Organization and planning**

<table>
<thead>
<tr>
<th>Weak</th>
<th>Less than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>More than satisfactory</th>
<th>Outstanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

**Decision making**

<table>
<thead>
<tr>
<th>Weak</th>
<th>Less than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Outstanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

**Leadership**

<table>
<thead>
<tr>
<th>Weak</th>
<th>Less than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Outstanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

**Perception**

<table>
<thead>
<tr>
<th>Weak</th>
<th>Less than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Outstanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

**Oral communications**

<table>
<thead>
<tr>
<th>Weak</th>
<th>Less than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Satisfactory</th>
<th>More than satisfactory</th>
<th>Outstanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>
REFERENCE LIST


Newcomb, T. An experiment designed to test the validity of a rating technique. Journal of Educational Psychology, 1931, 22, 279-289.

