Journal of Occupational Psychology (1990), 63, 331–342 *Printed in Great Britain*

Reliability of exercise ratings in the leaderless group discussion

Robert Gatewood*
University of Georgia, College of Business Administration, Department of Management, Brooks Hall 419, Athens, GA 30602, USA

George C. Thornton, III
Colorado State University

Harry W. Hennessey, Jr
University of Hawaii at Hilo

The demonstrated low correlations of dimension ratings across assessment centre exercises has led to reservations about their use. A proposed alternative is the use of overall exercise ratings that may represent total performance on work simulations. Three forms of reliability of such exercise ratings were studied in this laboratory experiment. The study design incorporated four groups of assessors and 31 participants in multiple Leadership Group Discussion (LGD) problems. (LGD was chosen because it is extensively used in assessment centres.) Inter-rater reliability, as measured by intra-class correlations within an assessor group, ranged from .69 to .99. Intergroup reliability, as measured by correlations of consensus ratings between assessor groups that had observed identical LGD groups, was .66 to .84. Alternate form reliability, as measured by correlations of overall ratings by the same assessor of participants in two different LGD problems, was .35 to .62. The implications of these findings for the use of overall exercise ratings are discussed.

In the traditional assessment centre (AC) method, dimensions serve as the basis for assessors' judgements. Assessors classify observed participants' behaviours into dimensions, share dimension observations (and sometimes dimension ratings) with other assessors, and rate overall performance on dimensions across exercises. Assessors may also integrate final dimension ratings into an overall assessment rating. However, recent research has questioned the psychometric adequacy of dimension ratings and has prompted the suggestion that behavioural observations should be summarized by ratings of overall performance in exercises rather than by overall dimension ratings. The purpose of this study was to examine the psychometric adequacy of such exercise ratings by measuring various forms of reliability.

* Requests for reprints.

Dimensions vs. exercise ratings

The design and execution of the traditional AC has been driven by dimensions. Dimensions are the employee characteristics that job analysis has shown to underlie effective job performance. Dimensions are used to direct the construction of situational exercises; they are the schema which guide assessors' observations, and they are the categories into which behavioural observations are classified. In the integration discussion, assessors typically share observations for the dimension categories, individually make overall dimension ratings, and then come to consensus about final dimension ratings. These final dimension ratings, in turn, are the basis of an overall assessment rating of suitability for selection. They can also serve as the basis for recommendations about training needs.

Recent research, however, has questioned the psychometric adequacy of dimension ratings. Factor analytic studies of final dimension ratings have typically yielded a small number of factors—occasionally a single factor attributed to halo and seldom more than three factors (Russell, 1985; Sackett & Hakel, 1979). Regression of the overall assessment ratings on final dimension ratings has shown that a small number of dimensions, typically three to five, account for a major portion of the variance in the overall assessment ratings (Dugan, 1987; Moses, 1972; Sackett & Hakel, 1979). Investigations of within-exercise dimension ratings have shown a lack of construct validity, specifically a lack of discriminant validity (Thomson, 1970; Turnage & Muchinsky, 1982). Factor analyses of within-exercise dimension ratings from several exercises have tended to yield exercise factors rather than dimension factors (Bycio, Alvares & Hahn, 1987; Robertson, Gratton & Sharpley, 1987; Sackett & Dreher, 1982). Also, Dugan (1988) found that increasing assessor training from two to three weeks did not diminish the occurrence of high intercorrelations among trait ratings. This lack of training effect could be due to cognitive limitations of assessors, including the possibility that what can be learned is learned in a relatively short period of time, the inherent redundant nature of dimension information, or a combination of these.

These results have suggested to some that assessors may not be able to assess performance as AC developers intended (Russell, 1985; Sackett & Dreher, 1982). It has been suggested that dimensions may not be the most appropriate organizing heuristics and that assessors should instead provide ratings of performance in exercises (Sackett & Dreher, 1982). Such exercise ratings could be interpreted as indicants of overall performance in testing situations closely representing job activities. The exercise would, therefore, be regarded as a 'job sample'. Traditionally, job sample tests have measured overall performance and not specific KSAs (behavioural dimensions in ACs). Such an explanation seems compatible with the conceptual strategy of ACs that exercises represent 'job simulations'. Before such modifications of the traditional AC method are instituted, however, research on the measurement properties of overall exercise ratings is needed.

Reliability of exercise ratings

As with any measurement device, a basic requirement of the use of exercise ratings is to demonstrate reliability. Only a few studies have investigated the reliability of overall ratings of effectiveness in exercises in an AC. All of these studies have focused on inter-rater reliability alone. For example, two studies reported inter-rater reliabilities for management games. Bray & Grant (1966) determined estimates of .60 for ratings and .69 for rankings. In Greenwood & McNamara (1967), the estimates were .74 for ratings and .75 for rankings, with subgroup estimates varying from .18 to .88. Bray & Grant's study was the only one which examined the in-basket exercise; reliability was estimated at .92.

The most often studied AC exercise has been the Leadership Group Discussion (LGD). The results of these studies have generally been consistent. Bass (1954) reported inter-rater reliability estimates of .61 to .84. Bray & Grant (1966) studied two different LGDs with estimates of .60 and .75 for ratings and .69 and .75 for rankings. In a similar study, Greenwood & McNamara's (1967) estimates ranged from .48 to .84 for 12 different groups. Finally, Clingenpeel (1979) reported inter-rater agreement of .72 and .69 for two LGDs.

Jones (1981) reported a study of four exercises frequently used in a UK military assessment centre. The study was prompted by Jones' questioning of the extent to which the pooling of judgements through discussion of assessors (a common procedure) affected reported inter-rater reliabilities. Referring to Huck's (1973) statement that such discussion may artificially increase reliability estimates, Jones investigated the level of inter-rater reliability before and after the pooling of judgements. He found that '. . . 42 to 53 percent of variance is typically common . . . before discussion . . . (and) . . . between 59 and 74 percent after . . .'. These differences in the magnitude of the common variance among raters pre-and post-discussion are important in that the increase after discussion could indicate the influence of social pressure to adjust ratings rather than agreement upon true score. Herriot, Chalmers & Wingrove (1985) addressed the question of social influence and concluded that it did have an influence on post-discussion ratings. Therefore, post-discussion ratings would demonstrate inflated inter-rater reliability. Because of this, the authors challenged the continued use of consensus discussion and the resulting ratings.

In his discussion, Jones states that there are several important reliability issues that should be examined in order fully to understand exercise ratings. Inter-rater reliability is the simplest issue, and the foundation for the others, but on its own does not provide complete information. Three other forms of reliability are alternate form reliability (where more than one version of a type of exercise is used), internal consistency reliability (where different exercises are used to generate information on a common number of dimensions), and repeat reliability (the use of the same exercise more than once).

Based upon Jones' remarks, the purpose of this study was to examine three aspects of the reliability of exercise ratings: inter-rater (before the pooling of judgements), alternate form (as defined by Jones), and intergroup (not mentioned by Jones but, as explained in the following paragraph, logically important).

Data about each of these forms of reliability are necessary for the serious consideration of restructuring ACs around ratings of overall performance in exercises. Inter-rater reliability is important because in a typical AC each participant is observed in a given exercise by only one assessor; it is important to know whether the ratings are consistent across assessors. As Jones (1981) points out, this form of reliability is basic and should be measured before discussion among assessors. Intergroup reliability is important because various assessor teams are used in an on-going AC. These teams can vary according to several characteristics, including AC experience, percentage of line management or human resources staff members, and functional area of members. The tacit assumption of ACs is that these groups are interchangeable because of training and have no independent effect on ratings. One way of examining this issue is to estimate the consistency of evaluations made by different assessor groups of the same exercise participants.

Alternate form reliability is also important given the manner in which ACs are conducted. Commonly, an organization conducting an AC possesses different forms of the same exercise, e.g. various LGD problems, management games, cases, presentation topics, etc. (The in-basket is often the only exception.) For a given AC session, one of these forms is selected for use. The implicit assumption is that these forms are interchangeable, and a participant should perform the same regardless of which problem he/she is assigned. If this is not the case, adjustments should be made in the training of assessors or the scoring of participants. In participant interaction exercises, like the LGD, the issue is compounded due to individual differences in the abilities of the other group members. The assumption, based on the common use of this exercise, is that group members, as well as the exercise problem, are interchangeable. If not, portions of each participant's rating would be attributable to differences in problem content and/or individual differences of other group members. Alternate form reliability has implications for: (*a*) ratings of participants made by assessors who were trained while using a different form/group combination of the exercise; (*b*) comparisons of participants in the same organization who attended different AC sessions; and (*c*) comparisons of participants in the same AC session who participated in exercises made up of different subgroups of participants.

The previously discussed, somewhat limited, research on the reliability of exercise ratings suggests that: (*a*) while inter-rater reliability averages in the .60s to .70s, estimates vary from the .20s to the .80s. Also, discussion among assessors may artificially increase inter-rater reliability; (*b*) agreement between groups of assessors seems to be lower than inter-rater reliability; and (*c*) no research addresses what Jones (1981) refers to as alternate form reliability.

Method

Overview

The LGD was chosen as the appropriate exercise for this study because it is used extensively in ACs, and it has been used in much of the previous research on the reliability of exercise ratings. To study the three types of reliability, the following characteristics were incorporated into the research design: (*a*) teams of assessors observed and rated participants in a LGD (inter-rater reliability), (*b*) two different teams of assessors observed and rated the participants of the same LGD (intergroup reliability), and (*c*) the same participants were observed and rated in two different LGDs by the same assessors (alternate form reliability). The design included both multiple groups of assessors and multiple LGD problems. Thirty-one discussants participated in two or three LGD problems. Four assessor groups were formed. One group observed all discussants in each of three LGD problems; each of the other three assessor groups observed two participant groups in only one LGD problem.

LGD problems and procedures

Three competitive LGD problems were used. Problem 1 required each participant to try to have a subordinate selected for a supervisory training programme. Thirty-one undergraduate students participated as discussants and were randomly assigned (within constraints imposed by class schedules) to six groups that varied in size from four to six. In problem 2, each discussant tried to obtain money from a limited pool for use on a particular project for the school system. Twenty-five undergraduates participated and were assigned to five groups of four to six. Assignment was made with the intention of minimizing overlap in group composition from problem 1. Problem 3 was a city council problem, similar in purpose to problem 2. Twenty-four students served as discussants and were assigned to four groups of six persons each, also with the intention of minimizing previous group overlap. Twenty-four students discussed problems 1 and 2, 23 students discussed problems 1 and 3, and 18 students discussed problems 2 and 3.

All problem discussions were conducted in the same room, around a circular table, with seating positions assigned by role. Each discussion was videotaped, using two cameras mounted high on opposite walls, so that each participant could be taped throughout the discussion. These videotapes were viewed by the various assessor groups for making evaluations. Each videotape of an LGD group was shown in its entirety only once. Assessments of videotaped interactions are not uncommon (Byham, 1986).

Subjects

Four groups of graduate students served as assessor subjects. Four PhD students made up group 1 (G1). These students had completed a graduate course in ACs, including practice in observing and evaluating behaviour in simulations. Groups 2, 3 and 4 (MBA 1, 2 and 3) consisted of 19 Master of Business Administration (MBA) students who were trained for eight hours in LGD observation and evaluation. The first two hours of the training were devoted to familiarizing the MBA assessors with the purpose and administration of LGDs as AC devices. Lecture and discussion covered the development of the AC concept, assessment dimensions, and the function of a LGD as a means of eliciting observable behaviour. The next two hours were devoted to training on the specific dimensions to be used in these LGD problems. Assessors were furnished dimension definitions and several behavioural examples. In addition, a behaviour-rating exercise was conducted in which the assessors were trained in the differences between judgement evaluations and behavioural observations. In the final four hours, each assessor participated as a discussant in an LGD group and practised observation and rating skills on groups composed of other MBA assessors. After training, MBA subjects were assigned to one of the three assessor groups.

Measures

Two measures were used for analyses in this study: (*a*) an individual assessor's rating (without discussion) of overall effectiveness in the exercise, made on a five-point scale, and (*b*) a calculated consensus rating among assessors, also based on a five-point scale. In order to rate the LGD participants, assessors watched the videotapes only once and made notes on behaviours. Each assessor rated discussants on 15 behavioural dimensions and then arrived at his/her own overall assessment rating of each discussant, using a five-point scale. The same type of rating procedure was reported by Russell (1987) in his study of the correlation of person characteristics and role congruency with exercising ratings.

The independent ratings of the overall effectiveness of each discussant were used to study inter-rater and alternate form reliability. The second measure, the calculated consensus rating, was derived for each LGD participant and used to study intergroup reliability. This measure was obtained using the procedures developed by Sackett & Wilson (1982). If 75 per cent or more of the assessors agreed in their independent overall rating of a discussant, this modal rating was used as the calculated consensus rating. If less than 75 per cent agreement occurred, the mean of the ratings was rounded and used as the consensus rating. Sackett & Wilson provided evidence that this procedure can reproduce the results of the traditional consensus-seeking process of arriving at overall ratings with a 94.5 per cent accuracy rate.

Design

Figure 1 shows the groups of LGD participants and assessors in the study. The G1s viewed and rated participants of all groups in all three LGD problems, a total of 15 groups. Each group of MBA assessors viewed and rated participants in two of the four groups of LGD problem 3. All three MBA groups viewed and rated discussion group 1; then, each MBA group viewed and rated one of the three remaining discussion groups.

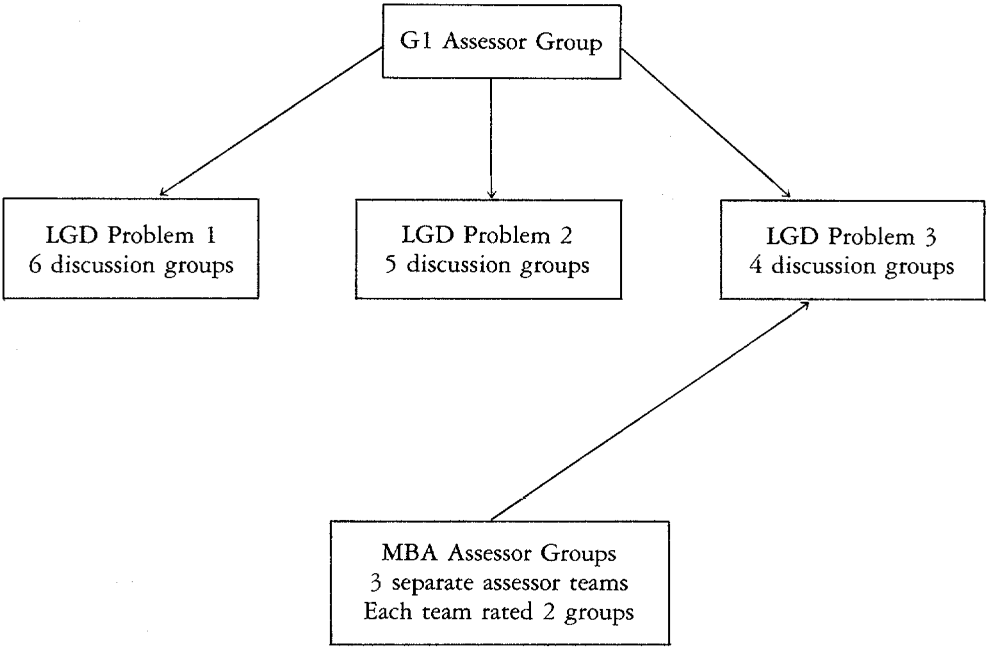

Figure 1. Schematic plan of the subject groups and events of the study.

Results

To test inter-rater reliability, intra-class correlations were computed for each assessor group for each LGD group observed. That is, 21 such correlations were produced: 15 for the G1 group and six for the various MBA groups. The independent overall ratings provided by each assessor for each LGD discussant were used in these calculations. Table 1 presents these correlations. All correlations indicate high inter-rater reliability, with coefficients ranging from .69 to .99 with a median of .93. Only two of these are below .80, five are between .80 and .89, and 14 are greater than .90.

Because frequently only one assessor observes and rates behaviour for an exercise, estimates of reliability of a single rater were computed (Guilford, 1954) and are shown in parentheses in Table 1. As expected, these reliabilities are considerably lower; they range from .43 to .95, with a median of .70. Ten of these reliabilities are below .70, nine between .70 and .89, and two above .90. These results indicate acceptable levels of reliability, even using ratings obtained without assessor group discussion.

Alternate form reliability was also examined at both the individual and the group levels. In each of these examinations the ratings supplied by only the G1s were used. Recall that the G1 group was the only assessor group to observe the same discussants in more than one LGD.

The independent overall ratings made by the G1s were used in the individual analysis. Correlation coefficients were computed for each G1, for each combination of two LGD problems. For example, 24 discussants participated in both LGD problems 1 and 2. Each

Table 1. Intra-class correlations within assessor groups

G1 assessors			
		LGD problem	
Discussion group	No. 1 (4 raters)	No. 2 (4 raters)	No. 3 (3 raters)
1	.95 (.81)*	.96 (.85)	.69 (.43)
2	.98 (.68)	.90 (.69)	.95 (.86)
3	.84 (.57)	.94 (.79)	.76 (.52)
4	.88 (.64)	.88 (.64)	.91 (.78)
5	.97 (.88)	.98 (.91)	
6	.80 (.50)		
MBA assessors			
LGD problem no. 3			
Discussion group	No. 1 (7 raters)	No. 2 (6 raters)	No. 3 (6 raters)
1	.91 (.58)	.97 (.85)	.93 (.70)
2	.95 (.74)		
3		.87 (.52)	
4			.99 (.95)

* Correlations in parentheses are estimates of reliability for one rater.

G1 provided an independent overall rating of each of these 24 discussants for each of the two LGD problems. The correlation coefficient was calculated for these two sets of ratings.

Results are presented in Table 2. The correlations are generally low, ranging from .35 to .62. Only four of 10 correlations are statistically significant, and even these are not of a magnitude generally thought to be acceptable for reliability estimates. It should be noted that these are reliabilities for a single rater. As mentioned previously, sources of variation include the members of the discussion group and the content of the LGD problem.

Table 2. Correlations of G1 ratings of individuals in two LGD problems

	LGD problem combination		
G1 assessor	1–2 ($N = 24$)	1–3 ($N = 23$)	2–3 ($N = 18$)
1	.39	.61**	.44
2	.35	.62**	.35
3	.45*	.55**	.43
4[a]	.36		

$*p < .05$; $**p < .01$.

[a] This assessor only observed LGD problems 1 and 2.

As a second test of the reliability of the ratings made by an individual assessor across alternate forms, comparisons were made of the differences in the magnitude of the ratings used to calculate the correlations. For each G1, the two ratings given the same discussant in two LGD problems were compared. The frequency of differences between the two ratings across the three LGD problem combinations was obtained. Table 3 contains this information. Of 219 pairs of ratings, 72 (33 per cent) were identical and 96 (44 per cent) were within one rating point of one another. In other words, in 51 (23 per cent) of the comparisons did the discussant receive ratings in two different exercises which differed by more than one point. All ratings were made on five-point scales. It should also be noted that there were approximately the same number of increases as decreases in ratings between the two LGD problems. Also, no ordering effect was found, i.e. no differences in mean ratings between the problems.

Table 3. Frequency of agreement of ratings of same individual in two different LGD problems

	G1 assessor				
	1	2	3	4	Total
Agreement	18	25	23	6	72
1 point difference	33	22	27	14	96
2 points difference	13	14	11	2	41
3 points difference	1	4	4	1	10
					219

Analysis of alternate form reliability at the assessor group level used the calculated consensus ratings of the G1 assessor group. Correlations were computed for the discussants who participated in each pair of LGD problems. For example, for the 24 discussants who participated in LGD problems 1 and 2, a correlation was computed using the calculated consensus ratings assigned to the discussant in each problem. These correlations provide data as to the reliability of group ratings given to an individual across two different LGD problems. The correlations were: .45 (LGD problems 1 and 2), .60 (LGD problems 1 and 3), and .40 (LGD problems 2 and 3). Only the first two correlations were significant ($p < .05$); however, even these did not demonstrate acceptable levels of reliability.

Results of these analyses offered, at best, only partial support for alternate form reliability. Correlational analyses indicated very low reliability for both individual and group ratings of individual discussants participating in two different LGD problems. However, the investigation of the differences in magnitude between the ratings of individual assessors indicates substantial agreement if a difference of one scale point is regarded as not being significant. If one were to judge reliability as being indicated by perfect agreement, however, then alternate form reliability would not be supported by any of the analyses.

To study intergroup reliability, the calculated consensus ratings were used. Ratings of performance of the individuals in the four discussion groups in problem 3 were used in these analyses because these were the only discussion groups that were observed by more than one assessor group. Correlations were computed on the calculated consensus ratings between the G1 group and those of each of the three MBA assessor groups for 12 participants in the two discussion groups observed by both assessor groups. As depicted in Fig. 1, the G1 group viewed each of the four discussant groups in LGD problem 3, while each MBA group viewed discussion group one plus only one of the other three discussion groups. Therefore, each MBA group observed two discussion groups (12 discussants) that were also evaluated by the G1 group.

The ratings of the G1s correlated .81 with MBA No. 1, .84 with MBA No. 2, and .66 with MBA No.3 ($p < .01$ for each). Similar to the analysis conducted for alternate form reliability, agreement on the magnitude of the calculated consensus ratings was also examined. Of the 36 pairs of ratings, 13 were in perfect agreement, 22 had one scale point difference, and the remaining pair had a two scale point difference. The results of these analyses were judged to support intergroup reliability, especially if one scale point difference is not regarded as being significant.

Discussion

The general conclusion that can be drawn from this study is that the reliability of assessors' overall exercise ratings in the LGD depends upon which form of reliability is being considered. As Jones (1981) has commented, in order to understand AC exercise ratings it is necessary to understand the various reliability types. The different forms of reliability in exercise ratings are not interchangeable.

Concerning the first of the three types of reliability studied, inter-rater reliability within assessor groups was generally high, ranging from .69 to .99 for four different assessor groups who rated a total of 21 LGD groups. This finding agrees with the results from previously cited studies. Two conclusions can be made from these results. First, relative to Jones' (1981) observation that inter-rater reliability is lower before assessor group discussion than afterward, in some cases training can apparently lead to substantial agreement among assessors even before discussion. Second, the similarity between the inter-rater reliability estimates of this study and those reported in the previously cited studies supports the generalizability of the results of this study.

The magnitude of these inter-rater reliability estimates is much higher than the reliability estimates for dimension ratings in previously summarized studies. This may indicate that overall exercise ratings are more consistent among members of an assessor group than are trait ratings across exercises and assessors. In terms of importance in ACs, as Jones pointed out, inter-rater reliability is basic. If assessors within groups do not agree, any resultant data would not be useful.

If the reliability of a single rater within an assessor group is considered, the estimates drop markedly. However, the median estimate for a single rater is .69, a value still much larger than the typical reliability of dimension ratings across exercises reported in previous studies (Robertson *et al.*, 1987; Sackett & Dreher, 1982).

There is also support for the reliability of overall consensus ratings of individual discussants between two different assessor groups (intergroup reliability). This indicates that training does convey a similar set of standards of behaviour to various groups of assessors. Obviously, this type of reliability is important in any consideration of the use of overall exercise ratings because assessor groups change frequently in the normal operation of an AC.

The most problematic type of reliability for LGDs is the third type, alternate form. In the present study, the reliability of ratings of individuals in two different LGD problems was much lower than the other two reliability forms. As noted previously, the content of the LGD problem and the other discussants in the LGD vary in this assessment of reliability. The common method of administration of ACs, however, ignores these variations. LGD participants usually are placed within groups and assigned a specific problem. There is no evaluation of differences in personal characteristics of other discussion group members or situational characteristics of the problem. The apparent assumption is made that all LGD groups are parallel forms.

It was beyond the scope of this study precisely to isolate the causes of the unreliability, but some inferences seem warranted. Individual differences among the discussants involved in a particular discussion group would seem to be a major source of unreliability. The behaviour of an individual discussant is directly influenced by the behaviour of other group members. Since it is not possible to standardize the behaviour of other group members, individual differences in abilities among other discussants would seemingly contribute to differences in the overall rating of the individual being assessed. Lack of standardization in the group composition is, therefore, thought to be a major cause of unreliability. Additional research is needed on the reliability of exercise ratings in non-interactive exercises such as the in-basket, or an oral presentation, to more fully explain this issue.

A second potential cause of unreliability is that differences in ratings of the same individual in two discussion groups may be due to dissimilarity of LGD problems. Table 2 indicates higher correlations by all three G1s for participants in LGD problems 1 and 3 than for the other problem combinations. In fact, all of these reliability coefficients are statistically significant, even though still quite low as reliability estimates. Similarly, the correlation for the calculated consensus ratings for participants in LGD problems 1 and 3, .60, was also significant and higher than the correlations for the other two LGD problem combinations. The pattern leads us to believe that similarity in some exercise characteristics, at this time unknown, underlies these findings.

These low correlations for individuals across discussion groups have direct implications for the previously discussed low correlations found for dimension ratings across different AC exercises. If the topic of the exercise and the mix of participants detracts substantially from the reliability of ratings made by the *same* assessor on the *same* participants in two forms of the *same* exercise, the expectation of high reliability of dimension ratings across exercises and assessors is unwarranted. Dimension ratings are, typically, incompletely nested within exercises. However, they are regarded as instances of the same variable regardless of the exercises in which they are incorporated. The findings of low alternate form reliability argue against such use of dimensions and lend little support to the idea that dimension ratings can, in fact, be gathered as currently intended in ACs. That is, if ratings of the same individuals in two forms of the same exercise are unreliable, dimension ratings will demonstrate less reliability and convergent validity than intended because of the added variability in situational factors that affect these ratings.

Conclusion

The results of the study indicate consistency of ratings of LGD participants both within assessor groups and between different assessor groups. This consistency, obviously, supports the use of exercise ratings. However, stability in ratings of the same participant in different LGD exercises was relatively low. This instability is crucial because it implies that the exercise score of a given participant could be affected by the nature of the LGD problem and the characteristics of other LGD participants. Unfortunately, such factors are not taken into account in group assignment procedures or scoring methods; neither is it a factor that can be necessarily counterbalanced by the viewpoints of other assessors within the group.

If one is heartened by the high estimates of inter-rater reliability and the high levels of inter-group reliability, one might be led to endorse the suggestion to use overall ratings of exercise performance. On the other hand, the moderate estimates of single rater reliability and the evidence suggesting situation specificity of overall ratings should be a warning that judgements of overall effectiveness in exercises may have limitations similar to those of judgements about performance on dimensions.

More research is needed to evaluate the various forms of reliability of judgements of overall performance in other types of exercises, e.g. in-basket, simulations of one-to-one interactions, and presentations. Even more important is the need to study the construct and predictive validity of ratings of overall performance in exercises. Before the traditional method of designing ACs and structuring assessors' judgements is altered, more research on the reliability and validity of overall exercise ratings is needed.

Acknowledgements

The authors wish to thank two anonymous reviewers for specific comments that were incorporated into the revised manuscript.

References

Bass, B. (1954). The leaderless group discussion. *Psychological Bulletin,* 51, 467–492.

Bray, D. W. & Grant, D. L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs,* 80 (17), Whole no. 625.

Bycio, P., Alvares, K. M. & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology,* 72, 463–474.

Byham, W. C. (1986). The assessment center method and methodology: New directions and technologies. *Monograph VII.* Pittsburgh: Development Dimensions International.

Clingenpeel, R. (1979). Validity and dynamics of a foreman selection process. Paper presented at the Seventh International Congress on the Assessment Center Method, New Orleans, LA.

Dugan, B. A. (1987). The use of information in assigning an overall rating: Can assessor training make a difference? Paper presented at the Fifteenth International Congress on the Assessment Center Method, Boston, MA.

Dugan, B. A. (1988). Effects of assessor training on information use. *Journal of Applied Psychology,* 73, 743–748.

Greenwood, J. M. & McNamara, W. J. (1967). Inter-rater reliability in situational tests. *Journal of Applied Psychology,* 51, 101–106.

Guilford, J. P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

Herriot, P., Chalmers, C. & Wingrove, J. (1985). Group decision making in an assessment centre. *Journal of Occupational Psychology,* 58, 309–312.

Huck, J. R. (1973). Assessment centers: A review of the external and internal validities. *Personal Psychology,* 26, 191–212.

Jones, A. (1981). Inter-rater reliability in the assessment of groups exercises at a UK assessment centre. *Journal of Occupational Psychology,* 54, 79–86.

Moses, J. L. (1972). Assessment center performance and management progress. *Studies in Personnel Psychology,* 1, 7–12.

Robertson, I., Gratton, L. & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology,* 60, 187–195.

Russell, C. J. (1985). Individual decision process in an assessment center *Journal of Applied Psychology,* 70, 737–746.

Russell, C. J. (1987). Person characteristic versus role congruency explanations for assessment center ratings. *Academy of Management Journal,* 30, 817–826.

Sackett, P. R. & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology,* 67, 401–410.

Sackett, P. R. & Hakel, M. D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. *Organizational Behavior and Human Performance,* 23, 120–137.

Sackett, P. R. & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology,* 67, 10–17.

Thomson, H. A. (1970). Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. *Journal of Applied Psychology,* 54, 496–502.

Turnage, J. J. & Muchinsky, P. M. (1982). Transituation variability in human performance with assessment centers. *Organizational Behavior and Human Performance,* 30, 174–200.

Received 9 November 1989; revised version received 6 June 1990