

Number and Type of Rating Dimensions on the Quality of Selection Interview Decisions

Mark A. Johnson & James P. Jolly
College of Business
Idaho State University

The effects of the number of rating dimensions on the quality of selection interview decisions were investigated. Two rating instruments were used--one with four performance dimensions and the other with nine dimensions. A number of lines of research suggest that fewer (e.g., three to four) rating dimensions result in superior quality ratings than ratings requiring evaluation on more (e.g., seven to nine or more) rating dimensions (Gatewood & Feild, 1990; Gaugler & Thornton, 1989). In the present study, the ratings of subjects under the four-dimension condition were hypothesized to have higher rating quality than the decisions of raters under the nine dimension condition. Results indicate that although psychometric rating properties did not differ significantly between the two treatment conditions, important practical differences in selection outcomes resulted.

The deficiencies of the selection interview in terms of reliability and validity have been documented over most of this century (Arvey & Campion, 1982; Gatewood & Feild, 1990). More recent meta-analyses, however, have presented the effectiveness of the interview in a far more favorable light (Wiesner & Cronshaw, 1988). Apparently, practitioners and researchers alike are learning from the accumulated research evidence which points to a number of practices that may increase the reliability and validity of the interview. For example, systematic collection and evaluation of interview data is generally recognized as an effective approach (Wiesner & Cronshaw, 1988; Arvey & Campion, 1982). Systematic evaluation requires formal evaluation forms which usually contain a number of job-related dimensions or characteristics, along with a rating

scale for each specific characteristic and the overall evaluation.

The quality of subjective ratings has long been a concern of all personnel/human resource managers and affects decisions in a variety of functional areas such as performance appraisal, job analysis, job evaluation, and selection interviewing. Despite the apparent diversity of these activities, a number of common rating issues are applicable across these functions. For example, questions regarding how many dimensions and what specific dimensions should be included for rating are two important issues. This paper attempts to gain some insight into how many dimensions might be appropriate in a selection interview situation.

Most rating experts agree that job performance is multi-dimensional (Cascio, 1991). Therefore, an important rating issue concerns the number of candidate characteristics or dimensions that interviewers should attempt to measure. In general, the accumulated research evidence suggests that raters limit their ratings to a maximum of nine rating dimensions (Gaugler & Thornton, 1989; Hinrich & Haanpera, 1976; Russell, 1985). For example, factor analytic research has demonstrated that even when raters are instructed to consider all information and all dimensions, only a few dimensions are typically used (Hinrich & Haanpera, 1976; Russell, 1985; Sackett & Hakel, 1979; Schmitt, 1977). Similarly, the results from a number of studies using regression analysis demonstrate that when overall ratings are regressed on dimension ratings the results converge in the conclusion that assessors use only three to six dimensions in structuring their evaluations (e.g., Gaugler & Thornton, 1989).

Gaugler and Thornton (1989) found that subjects who rated a small number of dimensions classified behaviors more accurately and made more accurate ratings than subjects who rated a large number of dimensions. Similarly, Henderson (1989) argues that only a restricted number of distinctive job characteristics should be used in the identification, observation, and measurement of job worth for compensation purposes.

These diverse lines of research in other personnel rating areas may provide useful information for selection interview settings. Based on these research findings, we might conclude that rating quality is likely to decrease as the number of rating dimensions increases (from three to nine dimensions). However, the rating requirements involved in assessment center and job evaluation contexts may be different from those requirements involved in selection interview decisions.

The conclusions that may be drawn from a review of the rating re-

search specific to the selection interview context are more restrictive. This research suggests that at most four rating characteristics should be evaluated and that more dimensions are not better than fewer. According to Gatewood & Feild (1990), the validity of the interview is improved when a limited number of specific applicant characteristics are measured. But what dimensions should be assessed?

What Specific Dimensions Should Be Rated?

A second important issue is which specific characteristics are best assessed in the interview? Three reviews of the literature (Gatewood & Feild, 1990; Arvey & Campion, 1982; Ulrich & Trumbo, 1965) suggest that there are only a few types of characteristics or dimensions that can be accurately evaluated with the interview. These include: interpersonal relations, such as sociability; verbal communications, such as verbal fluency; and work motivation, such as dependability, conscientiousness, stability, and perseverance. It appears that the interview is particularly well suited to measure these characteristics compared to other assessment methods. In addition, the interview has been found effective for evaluating the analytical/decision making ability (mental ability) of candidates, although an applicant's mental ability may be more efficiently assessed with a short test (Gatewood & Feild, 1990). However, in practice only a moderate percentage (39 percent) of employers use general mental ability tests (Mathis & Jackson, 1991), whereas, the most common selection method may be the interview (Dipboye, 1992; Gatewood & Feild, 1990; Bureau of National Affairs, 1988). In the past, employers have assessed analytical ability during the interview and likely will continue to do the same in the future.

Based on this research and practice, four core performance rating dimensions were selected for inclusion on the rating forms in this study. These included (1) verbal communications, (2) interpersonal relations, (3) work motivation, and (4) analytical/decision making ability.

Should Dimensions Be Generic or Job Related?

Another question that arises is whether all interviews should attempt to evaluate these same four dimensions. Should not the characteristics or dimensions rated be based on job analysis? Surely, each of these four factors is not important to all jobs. Does the manual laborer who digs

holes in the field need high verbal fluency? Or high interpersonal skills? Or high mental ability? It seems a distinction must be made between the ability of the interview to reliably measure these characteristics and the importance and relevance (validity) of these characteristics in a given job context (Dessler, 1991).

These issues were addressed by Osburn, Timmreck, and Bigby (1981) who showed that interviewers who rated on specific and relevant dimensions accurately discriminated between the more qualified and less qualified applicants. Subjects who rated on general job dimensions were not able to accurately discriminate between applicants. In addition, there was greater rater agreement among interviewers who rated on the specific job dimensions. Similarly, Dessler (1991) argues that interviewers should focus on traits that are not only more reliably assessed during the interview, but they should focus on those traits with the greatest evidence of validity. Thus there seems to be little disagreement among researchers that job related characteristics rather than non job related factors are more effectively evaluated during the interview (Arvey & Campion, 1982; Schmitt, 1977; Wicsner & Cronshaw, 1988).

Extraneous Factors Assessed During the Interview

A very real problem in ratings is the tendency for raters to inject into their ratings factors that are not part of the rating form and not intended to be rated. According to Gatewood and Feild (1990), there are many factors only marginally related to job activities that often influence even experienced interviewers' evaluations--these tendencies likely result in contaminated ratings. What if anything can be done to keep ratings pure? Osburn, Timmreck, and Bigby (1981) suggest keeping the questions and rating form dimensions job relevant. Another approach may be to have the raters evaluate these extraneous factors on specific dimensions, but then statistically remove the effects of these factors.

Hypothesis

This paper addresses issues concerning the number of ratings dimensions in the selection interview context. What effects, if any, does increasing the number of rating dimensions have on the reliability and accuracy of overall evaluations? We hypothesized that when the number of rating dimensions increases from four to nine the convergent and discriminant

validities will be lower, halo error will be higher (and halo accuracy will decrease), the reliabilities of ratings will be lower, and the overall accuracy of ratings will decrease.

METHOD

Subjects

Subjects were 234 upper-division students enrolled in human resource courses at a northwestern university. The average age of the subjects was 30 years. This student sample was selected because it was believed that this group, due to the relatively large proportion of older, nontraditional students, more closely resembles in age and background, "real life" interviewers. One hundred and sixteen subjects participated in the four-dimension treatment group and one hundred and eighteen subjects comprised the nine-dimension treatment group.

Rating Forms

The core performance dimensions comprising the four-dimension treatment condition were (1) verbal communication, (2) interpersonal relations, (3) work motivation, and (4) analytic/decision making ability. These are essentially the same set of dimensions recommended by Gatewood and Feild (1990). Each of these was defined on an accompanying form. The four core dimensions were selected based on the findings of prior work regarding the predictive effectiveness of these dimensions. To develop the nine-dimension rating form for the second treatment, five additional dimensions were added to the four dimension rating form. (See Appendix A for the nine dimensions and their definitions.) These dimensions were selected based on two sources. First, subjects in an earlier study (Johnson, 1990) who rated a subset of videotaped interviews were asked to list dimensions in addition to the four core dimensions they thought were important when evaluating the videotaped interviews. Second, research suggests job knowledge, education, and personal/professional appearance are among factors often found to affect interview ratings (Gatewood & Feild, 1990). Accordingly, the five additional factors were: job knowledge, education, professional appearance, fitness for the position, and potential for career advancement.

Interviews

Five videotaped interviews were used as stimuli for subject ratings. The five interviews were developed to allow for variability in dimension ratings among candidates. The variations in performances were scripted in such a way as to inject within-candidate variation across dimensions for at least a few of the candidates. An effort was made to balance the set of tapes so that the differences in interviewee quality were not so extreme as to make obvious the correct rank ordering of the interviewees on the performance dimensions or overall summary dimension. For example, the best candidate was not intended to be highest on all performance dimensions nor was the lowest quality candidate intended to be the worst on each dimension. Instead, tapes in this study were created to provide a mix of high and moderate qualities for the best candidate, and some low and moderate qualities for the poorer candidates. Thus, the videotaped interviews were developed in as realistic a manner as possible. During the debriefing sessions, many students indicated they were convinced the interviews were real.

Interviewees

Interviewees were graduate students in management. All of the interviewees were dressed in attire appropriate for the assistant personnel management position for which they were being interviewed. Interviewees were provided scripts for their roles. Practice sessions were held between the interviewees and the interviewer.

Procedure

Subjects in the four and nine rating dimension groups were shown five videotapes. They were shown each videotape and asked to rate the candidate at the conclusion of each viewing. Subjects in the four-rating condition made their ratings on the four core dimensions and an overall evaluation, whereas the subjects in the nine rating condition made their ratings on the four rating dimensions, the five additional dimensions, and an overall evaluation.

Development of Expert True Scores for Accuracy

Accuracy measures require the use of true scores in order to compare observed ratings. Estimates of true scores were obtained from expert ratings. In the present study, a Borman (1978, 1979) type approach was used to estimate the true performance level for each videotaped interview. Experts were ten doctoral students. The average of the expert ratings were used as true scores (average interrater reliability was .78, Spearman-Brown corrected reliability was .97). Accuracy was assessed by comparing the two subject groups' ratings to the experts' ratings.

DATA ANALYSIS

Convergent and Discriminant Validities

Convergent and discriminant validities for the four and nine dimension groups' ratings were obtained from a Ratee x Rater x Dimension ANOVA (Borman, 1978; Kavanagh, MacKinney, & Wolins, 1971). According to Borman (1978), a significant Ratee effect, especially one that explains a sizable proportion of the rating variance (Saal, *et al.*, 1980), supports convergent validity, and a Ratee x Dimension interaction indicates discriminant validity.

Reliability Measures, Halo Error and Halo Accuracy

Estimates of interrater reliability were obtained for each of the three rating groups (the 4- and 9- dimension groups and the expert group) by correlating each individual's ratings with the mean rating of the other raters. This was performed separately for each dimension. The average of these individual interrater reliabilities for each dimension was used as the reliability estimate for that rating dimension.

The dimension intercorrelation method was used to measure halo error. In addition, the average of the differences of the intercorrelations between the ratings of the experts and those of the subject group were used as measures of illusory halo for that subject group. The group with lower illusory halo has higher halo accuracy (true halo).

Accuracy

Overall accuracy was assessed by obtaining difference scores (each subject's overall ratings were compared to the mean of the expert group's

overall ratings). The difference scores on the overall ratings of the four- and nine-dimension groups were tested using Multivariate Analysis of Variance.

RESULTS

Convergent and Discriminant Validities

Convergent and discriminant validities for the four- and nine- dimension groups' ratings were obtained from Ratee x Rater x Dimension ANOVAs (Table 1). Results from both groups indicated significant convergent and discriminant validities (4-dimension group: convergent validity = .33, $p < .0001$, discriminant validity = .08, $p < .0001$; 9-dimension group: convergent validity = .29, $p < .0001$, discriminant validity = .06, $p < .0001$). For the 9-dimension group the convergent validity (.38) on the four core dimensions was higher than the convergent validity (.23) on the last five dimensions ($p < .10$).

Table 1

Convergent and Discriminant Validities of Ratings on Four Core Dimensions, and the 5 Extra Dimensions and All 9 Dimensions for the Nine-Dimension Group.

Group	Convergent Validity (Ratee Effect)	Discriminant Validity (Ratee*Dim)	Eta Square for the Model
Expert Group	.419	.209	.896
Four-Dimension Group	.326	.078	.842
Nine-Dimension Group			
4-Core Dimensions	.383	.060	.845
5-Extra Dimensions	.233	.036	.823
All 9 Dimensions	.288	.058	.806

Note: Sample sizes for the experts, four-dimension, and 9-dimension groups are 10, 116, and 118 respectively.

Interrater Reliability

The average intercorrelation interrater reliability for the four-dimension group was .65 and the average for the nine-dimension group was .62 but a *t*-test indicated that the difference was not statistically significant at the $p < .05$ level (Table 2). An additional analysis was conducted on the nine-dimension group in which the average interrater reliabilities on the four core dimensions were compared to the average on the additional five dimensions. The average correlation of ratings of the four core dimensions was .68 while the average for the additional five dimensions fell to .58 ($p < .01$).

Table 2

Average Interrater Reliability of Ratings Across Dimensions

	Average Reliability Across Dimensions
Expert Group	.78 ^a
Four-Dimension Group	.65
Nine-Dimension Group	
4-Core Dimensions	.68 ^b
5-Extra Dimensions	.58 ^b
All 9 Dimensions	.62

^a The average reliability for the expert groups's ratings is higher at a statistically significant ($p < .05$) level compared to the other average ratings provided in the table.

^b The average reliabilities for the nine-dimension group on the 4-core dimensions and the 5-extra dimensions had a statistically significant difference ($p < .05$).

Halo Error and Halo Accuracy

The average interdimensional correlations across the four core dimensions were .58 and .62 for the four- and nine-dimension groups, respectively (Table 3). Although the level of halo error for the four-dimension group was lower, the difference between the two groups (.04) was not statistically significant at the $p < .05$ level. Also, the average halo errors across the additional five dimensions and on all nine dimensions for the nine-dimension group were .52 and .59 respectively. These values were lower, not higher than the halo error from the four core dimension ratings made by the nine-dimension group.

Table 3

Halo Error for the Ratings of the Expert, Four-Dimension, and Nine-Dimension Groups

	Average Interdimensional Correlations (Halo)
Expert Group	.49
Four-Dimension Group	.58
Nine-Dimension Group	
4-Core Dimensions	.62
5-Extra Dimensions	.52
All 9 Dimensions	.59

Note: No statistically significant differences were found at the $p < .05$ level of significance

The average interdimension correlation for the expert ratings was .49. Thus, although the average halo level of the four-dimension group was more similar to the level of the expert ratings, this difference was not statistically significant at the $p < .05$ level. Therefore, the ratings of the

nine dimension group did not have greater halo error, nor did they have lower halo accuracy.

Overall Accuracy

The mean overall ratings given by the expert, four-dimension, and nine-dimension groups were 5.28, 5.96, and 6.38 respectively. Thus the overall ratings of the four-dimension group were closer to that of the experts' ratings than were the ratings provided by the nine-dimension group.

To statistically test the accuracy of the groups' ratings, the overall ratings given to each of the five ratees by each subject were compared to the mean overall ratings given by the experts for each of the corresponding ratees. This produced difference scores for each subject on each of the five ratees. The average of these difference scores for the four- and nine-dimension groups was compared. MANOVA results of the group effect for these difference scores on the overall ratings are provided in Table 4. The results indicated that the mean level difference scores for

Table 4

Multivariate and Univariate Analyses of Variance of Overall Difference Scores (Accuracy)

Comparison	DF	F	p <
Multivariate	5, 226	4.73	.000
Univariate			
Overall for Ratee 1	1, 230	5.14	.024
Overall for Ratee 2	1, 230	10.21	.002
Overall for Ratee 3	1, 230	6.03	.015
Overall for Ratee 4	1, 230	14.10	.000
Overall for Ratee 5	1, 230	2.37	.125

the four- and nine-dimension groups were statistically significant ($p < .001$). The difference scores for the four-dimension group were smaller (smaller difference scores indicate higher accuracy) than those of the nine-dimension group. The average difference scores for the four- and nine-dimension groups were 0.68 and 1.05 respectively. Follow-up tests using univariate ANOVAs of the difference scores on the overall ratings given to each ratee indicated statistically significant group effects ($p < .05$) on the ratings given to four of the five ratees (Table 4).

DISCUSSION

These results can be interpreted from three different perspectives; two psychometric and one practical. First, the psychometric properties (reliability, convergent and discriminant validities, and halo accuracy) did not differ across treatment groups--the psychometric properties of the four-dimension condition were not superior.

Second, within the nine-dimension group, differences were found when comparing the psychometric properties of the four core dimensions with those of the additional five dimensions. Measures of reliability and convergent validity were superior for the four core dimensions.

Third, differences in the ratings provided by the two groups have important practical implications, even though the psychometric differences were not great. Raters in the 9-dimension group provided consistently higher ratings (leniency) than those in the 4-dimension group or the expert group. This can be attributed especially to higher ratings given on the additional five dimensions. Moreover, the four- and nine-dimension groups' ratings resulted in different rank orderings of the five candidates which resulted in two different sets of selection decisions. The four-dimension group chose as the top candidate the one who was predetermined as best (by the researchers) and who was selected as the best candidate by the experts. The nine-dimension group chose this candidate as second best.

Raters often express a desire to include a relatively large number of rating dimensions, due, presumably to a perception that important characteristics are left untapped otherwise. This study demonstrates that the psychometric properties of ratings were not affected by inclusion of five additional dimensions--raters were able to make psychometrically sound discriminations among ratees. However, the additional dimensions may result in less accurate ratings--that is, selection of someone other than the

best candidate.

The answer to this apparent paradox may be explained by examining the concepts of reliability versus validity. Reliability reflects the consistency of ratings whereas validity addresses whether the inferences made with the measure are correct--that is, are the best candidates being selected. In the present case, ratings made by both the four- and the nine-dimension groups were of similar reliability, halo, and convergent and discriminant validities. The addition of five rating dimensions did not lower the average or overall psychometric properties of ratings for the nine-dimension group.

An issue that arose is whether the inclusion of the additional five factors lowers the validity of ratings. Clearly the five additional factors affected not only the overall and average ratings of the nine-dimension group's ratings, but the additional five dimensions evidently affected the ratings given on the first four dimensions rated by that group. Did these ratings reflect higher validity or not? The results showed that the overall accuracy of ratings (as compared to the experts' ratings) given by the four-dimension group were higher.

It appears that the job-relevance of rating dimensions is critical. Which of the four or nine dimensions are job related? Are only some of them important to job performance? If so, then non job-related dimensions should not be included. Or alternatively, what if some are more important than others: Should these dimensions be differentially weighted?

In conclusion, it appears that raters are able to rate equally well in a psychometric sense with four or nine rating dimensions: they are able to rate as reliably, with comparable levels of convergent and discriminant validities, and with similar halo levels. But it also seems that for ratings to be valid, they should be based on those dimensions that are job-related.

REFERENCES

- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, 35, 281-322.
- Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology*, 63, 135-144.

- Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, 64, 410-421.
- Bureau of National Affairs, Inc. (1988). *PPF Survey No. 146--Recruiting and Selection Procedures* (Washington, D.C.: Bureau of National Affairs, Inc.).
- Cascio, W. (1991). *Applied Psychology in Personnel Management*. 4th ed. Englewood Cliffs: Prentice-Hall.
- Cooper, W. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.
- Dessler, G. (1991). *Personnel/Human Resource Management*, 5th ed., Englewood Cliffs: Prentice Hall.
- Dipboye, R. L. (1992). *Selection Interviews: Process Perspectives*. Cincinnati: Southwestern.
- Gatewood, R. D., & Feild, H. S. (1990). *Human Resource Selection*, 2nd ed., Chicago: Dryden Press.
- Henderson, R. I. (1989). *Compensation Management: Rewarding Performance*, 5th ed. Englewood Cliffs: Prentice-Hall.
- Hinrichs, J. R., & Haanpera, S. (1976). Reliability of measurement in situational exercises: An assessment of the assessment center method. *Personnel Psychology*, 29, 31-40.
- Johnson, M. A. (1990). The effects of videotape review on selection interview ratings. Presented at the 31st annual meeting of the Western Academy of Management, March 1990, Salt Lake City, Utah.
- Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin*, 75, 34-49.
- Mathis, R. L. & Jackson, J. H. (1991). *Personnel/Human Resource Management*, 6th ed. St. Paul: West Publishing.
- Osburn, H. G., Timmreck, C., & Bigby, D. (1981). Effect of dimensional relevance on accuracy of simulated hiring decisions by employment interviewers. *Journal of Applied Psychology*, 66, 159-165.
- Russell, C. J. (1985). Individual decision processes in an assessment center, *Journal of Applied Psychology*, 70, 737-746.
- Saal, R. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428.
- 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.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. *Journal of Applied Psychology, 62*, 171-176.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.

Wiesner, W. H. & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275-290.

Dr. Mark A. Johnson
College of Business
Idaho State University
Pocatello, ID 83209
(208) 236-2155

APPENDIX A

DEFINITIONS OF DIMENSIONS

Verbal Communication: Ability to communicate effectively, one-on-one, in small groups, and in public speaking contexts. Fluency, "quickness on one's feet," clarity, organization of thought processes, and command of the language are all important.

Interpersonal Relations: Favorable bearing and demeanor. Ability to win the liking and respect of others; cooperative; ability to deal successfully with a broad range of people; friendly. Tunes in accurately to feelings, needs, and attitudes of others; understands the impact of one's own behavior on others.

Work Motivation: Demonstrated willingness to maintain high activity level. Compelling desire to succeed. Maintenance of realistically high standards. Willingness to persevere to successful completion, despite obstacles. Actively seeks out opportunities to make a contribution.

Analytic and Decision Making Ability: The ability to analyze problems in depth and make decisions in a methodological, systematic, and decisive manner. Ability to generate ideas and practical, sensible, realistic solutions to problems. Ability to juggle multiple projects simultaneously.

Knowledge of the Field: Acquired knowledge from formal education, training, experience or some combination of these. Possesses the knowledge, skills, and understanding of the field required to perform the job.

Educational Preparation: Earned a bachelors or masters degree. Degree(s) in relevant fields such as personnel/human resources management, industrial relations, and management. Other related fields such as general business, organizational behavior, and industrial psychology with relevant coursework are also acceptable.

Professional Appearance: Portrays oneself in professional manner. Neat and clean in appearance. Well dressed, appropriate attire. Appearance would enhance the image of the organization when the individual acts in a public relations role.

Fitness for this Position: Possesses the qualities, skills, abilities, and aptitudes appropriate for this job, and the ability or potential to apply these characteristics to the job. A good person/job fit.

Potential for Career Advancement: The capability to move up in the organization. Likely to develop and be qualified for higher level responsibilities within the personnel function or other management areas.