

Increases in Interrater Reliability of Situational Interviews as a Function of the Number of Benchmark Answers

Lisa M. Buchner
Duke Power Company

A study was conducted to determine whether the interrater reliability of a situational interview increased as the number of benchmark answers increased. Half of the subjects (raters) received the usual three-benchmark format and half received benchmarks at all five scale points. The results indicated that benchmarks at all five positions resulted in interrater reliability of .66 compared to the .44 found when the traditional three benchmarks were used.

Interviews have long been the most used and relied upon employee selection method, and it is estimated that over 90% of all businesses in the United States use the interview as their major selection tool (Beach, 1985). It is unfortunate then that research has consistently documented low levels of reliability and validity in interview situations (Arvey & Campion, 1982; Mayfield, 1964; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949).

Conceptually, it is not difficult to understand this lack of reliability and validity. Instead of questioning the applicant about job related knowledge and skills, the interviewer poses such questions as, "Where do you expect to be 10 years from now?" Even a good answer means little, for it not only is unrelated to performance, but chances are that interviewers will not agree on what constitutes a good answer.

Realizing that managers remain committed to the use of the interview despite such negative evidence, research in the past 10 years has sought to increase interview reliability through standardizing, or structuring the process. According to Weekly and Gier (1987), the most successful areas of research in terms of application of this structuring lay in the behaviorally based approaches. The most successful of these is the situational interview. Developed by Latham in 1980, situational interviews are based on the premise that a person's expressed behavioral intentions are related to subsequent behavior (Weekly & Gier, 1987).

In a situational interview, applicants are presented with a number of job related situations and asked what they would do in each case. Designed to identify behaviors critical to effective performance on the job (Latham & Saari, 1984), critical incidents are obtained from a job analysis and transformed into questions. Because the applicants' responses are given a score between 1 and 5 (1=poor, 5=outstanding), interviewers are supplied with examples of responses (called "benchmark answers") that would warrant a 1, 3, or 5 rating in order to guide them in the scoring process.

Compared to an unstructured interview's interrater reliability coefficient of .35 (Landy, 1985), interrater reliabilities of situational interviews have been shown to be between .76 and .84 (Latham et al., 1980, Weekly & Gier, 1987). It is assumed that the utilization of benchmark answers partly explains this vast difference (Latham et al., 1980).

Because the purpose of providing these anchors is to reduce the subjectivity involved in scoring, it is logical to assume that agreement between raters would increase as a function of such a reduction.

As was previously discussed, the current format of a situation interview utilizes a 5-point rating scale with benchmark answers only at the 1, 3, and 5 levels. If one accepts the assumption that benchmark answers increase interrater reliability, it follows that adding one or more benchmarks to *every* level would further increase interrater agreement.

The purpose of this study was to test such an assumption. More specifically, it was hypothesized that by assigning benchmark answers to each point level (1, 2, 3, 4, 5), the interrater reliability of a situational interview would be greater than an interview using only one benchmark at levels 1, 3, and 5.

Method

Subjects

All 39 participants (19 male, 20 female) were students of psychology at a medium sized university. Of these, 14 graduate and seven undergraduate students served as raters, while 18 lower level students served as interviewees.

Procedure

This study involved three steps: (1) development of a situational interview for the position of teller at a credit union, (2) examination of the interrater reliability of that interview using 5 or more benchmarks as opposed to the same version using only three, and (3) data analysis.

Step 1: Using the method developed by Latham et al. (1980), over 185 critical incidents were obtained from a previously conducted job analysis. In addition, four credit union employees (ranging from teller to branch manager) were interviewed, and another 27 critical incidents were generated. All 212 incidents were reviewed and categorized into 6 job dimensions. Two to three incidents from each dimension were converted into questions.

Once the questions were developed, 3 graduate students generated as many answers as possible for each question and rated them on a scale of 1 to 5. It was decided that 2 questions were essentially equal to other ones so they were removed, resulting in a final interview of 14 questions.

Step 2: A total of 18 interviews were conducted. While the author actually administered questions to the interviewees, answers were scored by 4 raters who were present during the interview. To assist them in scoring, each rater had a copy of the interview questions, along with corresponding benchmark answers. All four had identical questions, although the number of benchmark answers varied. Two raters had 3 anchors weighted at the 1, 3, and 5 levels, and the other two had the same benchmarks, plus anchors at levels 2 and 4. After each question was answered by the interviewee, raters recorded their score on a separate answer sheet.

Step 3: For each subject, scores given by the two raters using three benchmarks were correlated with each other and the same was done with the raters using five or more benchmarks.

Results and Discussion

The interrater reliabilities of the scoring methods using either three or five benchmark answers were calculated to be .44 and .66, respectively. These findings seem to support the hypothesis that adding more benchmark answers to a situational interview results in increased interrater reliability.

One possible explanation for these results is that the addition of more benchmarks reduces the amount of subjectivity in assigning point values to an interviewee's responses. Considering that many answers are not clearly bad (1), average (3), or outstanding (5), providing examples of "in between" answer would serve to reduce the amount of judgment that might otherwise be necessary with only three benchmarks.

Another reason for the higher reliability of interviews with five or more benchmarks may be a matter of probability. Simply put, the more benchmarks included, the higher the probability that an interviewee's response would exactly match one of the benchmarks. No judgment is needed in the determination of an appropriate score. Therefore, any variability among scores would be severely reduced (if not eliminated) because the correct score is literally "given" to the interviewer. It would seem then, that in a situation such as this, it makes very little difference in interrater reliability whether or not 5 or 3 benchmarks were used. Since their purpose is to provide guidance for the rater, the benchmarks are useless under conditions when answers exactly match one of the benchmarks listed.

According to data in this study, however, the number of benchmarks had a definite impact on interrater reliability even when responses to questions generally matched the benchmarks. It is interesting to note that, in theory, if every possible answer to a question were listed as a benchmark, interrater reliability would be near perfect. The next question is: How many benchmarks would be needed to cover the entire range of answers? Also, is it necessary to cover every possible answer, or is there a limit to the reliability obtained? At what point does adding benchmarks cease to increase reliability?

Implications for further research may include the examination of practice effects on the reliability of the situational interview. Logic tells us that the provision of benchmarks assists raters in determining an appropriate point value, but after a person has conducted several interviews and has

heard the entire range of possible answers, benchmarks would seem to be unnecessary. It would be interesting to measure at what point this may occur.

The results of this study are promising for the future of the situational interview. Although reliability and validity is already much higher than that of an unstructured interview, the evidence shows that adding a few more benchmarks will increase the interrater reliability even more, ultimately resulting in a more effective selection device.

References

- Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281-322.
- Beach, D. S. (1985). *Personnel: The management of people at work*. New York, Macmillan.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.
- Landy, F. J. (1985). *Psychology of work behavior*. Homewood, IL: Dorsey Press.
- Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.
- Latham, G. P., & Saari, L. M. (1984). Do people do what they say? *Journal of Applied Psychology, 69*, 569-573.
- Mayfield, E. C. (1964). The selection interview: Reevaluation of published research. *Personnel Psychology, 17*, 239-260.
- Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*, 79-101.
- Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.
- Wagner, R. (1949). The employment interview: A critical review. *Personnel Psychology, 2*, 17-46.
- Weekly, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72*, 484-487.