Journal of Applied Psychology
1980, Vol. 65, No. 4, 422-427

The Situational Interview

Gary P. Latham and Lise M. Saari
University of Washington

Elliott D. Pursell and Michael A. Campion
Weyerhaeuser Company, New Bern, North Carolina

The situational interview is based on a systematic job analysis known as the critical-incident technique. The incidents are turned into interview questions in which job applicants are asked to indicate how they would behave in a given situation. Each answer is rated independently by two or more interviewers on a 5-point Likert-type scale. To facilitate objective scoring, job experts develop behavioral statements that are used as benchmarks or illustrations of 1, 3, and 5 answers. In Studies 1 and 2, the interobserver reliability coefficients for situational interviews of hourly workers (n = 49) and foremen (n = 63) were .76 and .79, respectively. Similarly, the internal consistencies of the interview questions for the hourly workers and foremen were .71 and .67, respectively. The respective concurrent validity coefficients were .46 and .30. In Study 3, predictive validity coefficients of .39 and .33 were obtained with women and blacks, respectively. All of these values were significant at the .05 level.

The interview is used as a selection device by virtually every company in the United States. In fact, the *Wall Street Journal* (Lancaster, 1975) reported that a majority of companies have phased out pencil-and-paper tests and rely solely on the interview for making hiring decisions.

The widespread use of the interview in favor of tests has occurred despite the fact that the interview is considered as much a test by government agencies as is a standardized test of intelligence or any other decision-making process that affects an individual's employment status in an organization. Nevertheless, companies appear to believe that the probability of being investigated by a government agency for wrongdoing in the areas of selection, promotion, layoff, and termination is reduced if only the interview is used as the decision-making instrument. What makes the reliance on the interview for making selection decisions alarming is that the interview often lacks reliability and validity (Mayfield, 1964; Ulrich & Trumbo, 1965; Wagner, 1949).

One reason why the interview often lacks reliability is that interviewers seldom ask the same questions of different applicants. Moreover, when the same questions are asked, interviewers frequently disagree on the desirability or appropriateness of the interviewee's responses. Lack of reliability is a serious problem in that it can attenuate validity (Thorndike, 1949).

A theoretical approach on which valid interviews might be based is Locke's (1968) theory of goal setting. The underlying assumption of this theory is that intentions are related to behavior. If what people say correlates highly with what they do, the advantage of using the interview for making selection decisions is obvious. The interview would approximate a sample of actual job behavior, and the need for expensive written aptitude tests would be reduced, as would the cost of developing job simulation exercises (e.g., in baskets). However, a potential problem with the interview that is generally not a concern with aptitude tests or job

Studies 1 and 2 were conducted by the first two authors; Study 3 was conducted by the second two authors. The authors would like to thank Patricia Cain Smith and Benjamin Schneider for their comments on this article.

Requests for reprints should be sent to Gary P. Latham, Department of Management and Organization, DJ-10, University of Washington, Seattle, Washington 98195.

 0021-9010/80/6504-0422$00.75

simulations is the social desirability response or faking. Many interviewees can quickly discern from the wording of a question the answer the interviewer wants to hear.

In an attempt to overcome some of these issues, Maas (1965) developed an interview procedure based on Smith and Kendall's (1963) recommendations for developing behavioral expectation scales. In brief, Maas's procedure involved having interviewers who were familiar with the job in question brainstorm traits that should be exhibited by effective job incumbents. Examples of on-the-job behaviors were then written by the interviewers to illustrate a high, average, and low degree of each trait. A second group of judges, unaware of which examples were written for a given trait or level (high, average, or low) reallocated the examples into traits and levels. Only examples with high agreement as to trait and level were retained. The anchors or illustrations for each level were then reworded as "expectations."

Interviewers subsequently rated each job candidate on each trait by making analogies from the candidate's interview responses to the behavioral anchors on the appraisal instrument. In a study of college orientation counselors, the interobserver reliability was .58. In a second study the interobserver reliability was .69. These coefficients were significantly higher than those that were obtained using rating scales benchmarked with adjectives (e.g., very good).

There are several limitations to Maas's study. First, the criterion-related validity of the procedure is not known. Second, the emphasis in the interview was on traits. It is doubtful whether such an interview would satisfy requirements for content validity. Moreover, the brainstorming of traits may leave much to be desired from the standpoint of a systematic job analysis. Third, the manner in which the questions for assessing each trait were formulated was not reported. Fourth, despite the reallocation procedure, the interobserver coefficients obtained from the interviewers were not high by conventional standards for evaluating a test. This may not be surprising in light of the approach used for the job analysis and the emphasis that was put on traits. Moreover, it is questionable whether the anchored examples should have been worded in terms of expected on-the-job behavior rather than actual interviewee behavior indicative of subsequent job behavior. Finally, the interviewers were college students interviewing applicants for a college job. Thus, the generalizability of the results to industrial organizations is not known.

The present research differs from that reported by Maas in that we were concerned with the validity as well as the reliability of our interview technique. In addition, a systematic *job analysis* was used to develop the performance appraisal instrument as well as the selection interview. With regard to the latter, the job analysis information was used to develop the actual interview questions rather than to benchmark the answers. Job experts benchmarked answers for scoring an interviewee's responses in terms of comments that they had heard in interviews that they believed identified employees who subsequently became poor, average, or excellent performers on the job. Studies 1 and 2 report the results of two concurrent validity studies that were conducted in an industrial setting in the northwestern United States for both an entry-level job and a first-line supervisory position. Study 3 reports the results of a predictive validity study for entry-level workers in a company in the rural South.

Studies 1 and 2

Method

Sample. Study 1 was conducted on unionized hourly sawmill workers. Forty-nine of these individuals were randomly selected from 207 employed in a company facility. Of the 49 people interviewed, all were male and 44 were white. The mean age was 29.4 years (SD = 2.5).

The participants in Study 2 were 63 first-line foremen, all white males. Their mean age was 43.3 years (SD = 10.6), and the mean number of years they had worked on the job was 5.4 (SD = 4.5).

Procedure. A job analysis was conducted using the critical-incident technique (Flanagan, 1954). The results for the foremen have been reported in detail elsewhere (Latham, Fay, & Saari, 1979). In brief, four performance criteria or behavioral observation scales (BOS; Latham & Wexley, 1977) were developed. Each BOS contained from 4 to 13 behavioral items that were rated on a 5-point Likert-type scale.

The job analysis for the hourly workers yielded nine criteria or BOS. Each BOS contained from 2 to 12

behavioral items. Because these criteria were developed independently of the first two authors (Pursell, Note 1), the appraisal format differed from that used for evaluating the job performance of foremen. First, each behavioral item was rated on a 6- rather than a 5-point scale. Second, after rating an individual on all items defining a given criterion or BOS, the rater made a global rating on a 9-point scale as to the overall effectiveness of the individual on that criterion.

The situational interviews for the hourly workers and the foremen were developed by 3–5 company supervisory people. These were superintendents who had experience in interviewing and supervising both hourly workers and foremen. The superintendents examined the critical incidents collected in the job analysis. These incidents reflected job areas such as attendance, safety, interaction with peers, work habits, and so forth. Each superintendent picked one incident that he believed exemplified the criterion under consideration and turned the incident into a question. Each question was read aloud to the group. Through group consensus, one or at most two interview questions were selected. Restricting each criterion to one or two questions was necessitated by the 60-minute time limit that the company believed could be devoted to conducting one interview.

An example of a critical incident describing ineffective behavior of an hourly worker was:

> The employee was devoted to his family. He had only been married for 18 months. He used whatever excuse he could to stay home. One day the fellow's baby got a cold. His wife had a hangnail or something on her toe. He didn't come to work. He didn't even phone in.

This incident was rewritten by the superintendents in the form of the following question:

> Your spouse and two teenage children are sick in bed with a cold. There are no relatives or friends available to look in on them. Your shift starts in 3 hours. What would you do in this situation?

Each member of the group was then asked to independently benchmark a 5 answer, that is, "things you have actually heard said in an interview by people who subsequently were considered outstanding on the job"; a 1 answer, that is, "things that you have actually heard said in an interview by people who as a result got hired but turned out to be very poor performers"; and a 3 answer, that is, "answers that you have actually heard said in an interview by people who as a result got hired and turned out to be mediocre performers." For job experts who did not have extensive interviewing experience (e.g., line managers), these instructions were modified to read "think of people you know who are outstanding, poor, and mediocre on the job. How do you think they would respond to this question if they were being interviewed?"

Each person then read his answers to the other group members. After group discussion, consensus was reached on the answers to use as benchmarks. The three benchmarks for the above question were: I'd stay home—my spouse and family come first (1); I'd phone my supervisor and explain my situation (3); and Since they only have colds, I'd come to work (5).

The reallocation step used by Maas was not used in the present study due to time constraints. The situational interview for hourly workers contained 17 questions; the interview for foremen contained 10 questions. A concurrent validity study was then conducted on each interview.

To ensure the cooperation of the hourly union workers who were interviewed, we stressed that it was hoped that the results of the study would bring about the selection of employees who would do their fair share of the work and not become a burden to them. We also emphasized that the results would not affect them directly because their answers would not be placed in their personnel files; but again, it was stressed that the test results would indirectly affect them in the sense that they would be affected by the performance of "losers who were hired but seldom fired around here." Similar comments were given to the foremen; however, no assurance was given that their test results would not be examined closely by upper management. When questions concerning this issue were raised, the personnel administrator stated, "Come on, you guys are big boys; you know you wouldn't be going through all this if it didn't count." This statement was given deliberately, according to her, to simulate the test anxiety experienced by job applicants.

In conducting the interview, one person read the question and two or more interviewers recorded the answer. The interviewee was told that the question would be repeated on request.

Twenty superintendents scored the interviews. Code numbers rather than names were used so that the supervisors' scoring of the interview could not be biased by knowledge of the interviewee's identity. The superintendents worked in pairs. Each answer was scored independently and then through discussion one rating was agreed on. Both the independent ratings and the consensus ratings were recorded.

Prior to having supervisors appraise the job performance of the hourly workers, the second author generated 15–20 minutes of group discussion with them on ways to minimize rating errors such as contrast effects, halo, similar to me, and so forth. The superintendents who completed the BOS on foremen in Study 2 received an 8-hour in-depth training course for minimizing rating errors in observing and evaluating others. This training program has been described in detail elsewhere (see Latham, Wexley, & Pursell, 1975).

On completion of this discussion and training, both the supervisors (n = 8) who evaluated the job performance of hourly workers and the superintendents (n = 20) who evaluated the job performance of the foremen worked alone when completing the performance appraisal forms. None had knowledge of anyone's performance in the situational interview.

Results

The mean interjudge reliabilities of the independent ratings for both the hourly

worker and foreman interviews were significant, $r(15) = .76$, $p < .05$ and $r(8) = .79$, $p < .05$, respectively. The internal consistencies of the hourly worker and foremen interviews were also satisfactory ($\alpha = .71$, $p < .05$ and $\alpha = .67$, $p < .05$, respectively). Internal consistency was desirable in this instance because of the moderately high intercorrelations among the BOS ($M = .58$). The intercorrelations do not necessarily imply halo error because industries, like universities, strive for homogeneity by discharging individuals who perform poorly in one or more areas. Moreover, the criteria are logically related. For example, the criteria for evaluating foremen tap different aspects of supervisory behavior as opposed to skills that are logically unrelated (e.g., physical vs. cognitive abilities). Multidimensional criteria are necessary because the measures do not overlap one another completely, and more importantly they facilitate accountability and control by the organization and feedback and development for the individual.

The results of the hourly worker interview validation in Study 1 indicated that the interview scores correlated significantly with each of the nine performance criterion areas on the BOS, correlation coefficients ranging from .28 to .51 (ps $< .05$); the interview scores correlated significantly with the sum of the nine "overall (global) ratings" that followed each performance criterion, $r(47) = .50$, $p < .05$; and the interview scores correlated significantly with the total BOS scores, $r(47) = .46$, $p < .05$). Partialing out experience did not reduce these correlations significantly, the latter two correlations dropping to .46 and .41, respectively.

The validation results for the foreman situational interview in Study 2 indicated that the interview scores correlated significantly with three of the four BOS. These were $r = .28$ for Safety, $r = .35$ for Work Habits, and $r = .31$ for Organizational Commitment (all ps $< .05$). The interview scores did not correlate with performance on the criterion Interaction With Subordinates. The intercorrelations among the four BOS ranged from .52 to .79. The interview scores also correlated significantly with the composite BOS score, $r(61) = .30$, $p < .05$. When experience was partialed out, this correlation was reduced to $r(61) = .29$, $p < .05$.

Study 3

Although well-conducted concurrent studies can provide useful estimates of validity, there is a possibility that test scores may be affected by additional job knowledge, different motivation levels, or added maturity of job incumbents versus applicants. For this reason the predictive validity of the situational interview was determined in a third study. This study was conducted with an organization in the rural South that has a strong Affirmative Action policy. Therefore, an additional purpose of this study was to determine the effectiveness of the situational interview in selecting females and blacks.

Method

Sample. The situational interview was administered to 56 applicants for entry-level work in a pulp mill, all of whom were subsequently hired. Of this number, 30 were female and all were black. The mean ages of the females and blacks were 31.5 years ($SD = 8.9$) and 30 years ($SD = 6.9$), respectively.

Procedure. The procedures for developing both the interview questions and the performance appraisal instrument were identical to those described for foremen in Study 1. Ten situational questions were developed.

Results

The mean interobserver reliabilities ($df = 8$) of the ratings on the situational interview were .87 and .82 (all ps $< .05$) for blacks and females, respectively. Similarly, the internal consistency (Cronbach's alpha) of the situational interview was .70 for blacks and .78 for females (all ps $< .05$).

The employees' job performance was evaluated after they had been on the job for 12 months. None of the supervisors who evaluated the employees were aware of how well any employee had performed in the situational interview. Prior to making their evaluations, the supervisors received the same training for minimizing rating errors (Latham et al., 1975) used in Study 2. A composite job performance rating was calculated for each employee. The correlation between performance in the interview and perform-

ance on the job 12 months later was .39 for females and .33 for blacks ($ps < .05$).

General Discussion

The results of these three studies provide further support for the theoretical proposition (Locke, 1968) that intentions correlate with behavior. Previous support for this assumption has been confined primarily to the motivational literature (cf. Latham & Yukl, 1975). The present findings are particularly impressive in light of the low reliability and validity of other interview methods, and the comparability of the validity coefficient for job performance of supervisors with that which is typically reported for assessment centers ($r = .33$; Cohen, Moses, & Byham, Note 2). Assessment centers generally last a minimum of 1 entire day, and many are conducted for 3 days. A situational interview can be conducted within an hour. This is not to imply that a situational interview should be used in place of an assessment center or other selection tests. However, it is likely that including this technique in an assessment center or with other test batteries would significantly improve the validity of the selection process.

The effectiveness of the situational interview is readily explainable. First, the interview questions are derived from the results of a systematic job analysis. A representative sampling of job situations is incorporated in the interview questions. Thus, the content validity of the procedure appears to be satisfactory as judged by job experts.

Second, the face validity of the procedure is ensured by asking only job-related questions. This appears to increase the motivation of the interviewee to take the test seriously.

Third, focusing on the interviewers' experience with a wide range of interviewee responses, and choosing among these responses to develop a scoring key to anchor 1, 3, and 5 answers, may have increased the interobserver reliability and validity of the procedure. The instructions to interviewers emphasized that these benchmarks were only illustrations or aids for scoring an answer. Interviewers were to use their judgment of what constituted a 1, 2, 3, 4, or 5 answer. The similarity between the answers given by each interviewee and one of the three benchmarks to each question, however, was striking. In short, the job experts who developed the scoring key turned out to be truly expert in predicting almost exactly how people would respond to each interview question.

Fourth, both the selection and the performance appraisal instruments were based on overt employee behavior rather than traits or economic constructs. Traits are generally ambiguous and thus unpredictable. Economic constructs are frequently affected by factors over which the job performer has little or no control. For effective selection it is necessary to develop predictors that are not only realistic samples of behavior but are as similar to the criteria as possible (Wernimont & Campbell, 1968). An interview by nature can usually tap only behavioral intentions. The present research has shown that when the intentions measured are job-related they can serve as valid indicators of on-the-job behavior. Nevertheless, the generalizability of these results cannot be assumed; the effectiveness of a situational interview must be demonstrated through proper validation.

The situational interview might be improved by combining it with an approach used by Ghiselli (1966), which focuses on *past* behavior rather than *future* intentions. Job candidates could be asked what they have done in the *past* in situations similar to those posed by the interviewer. Such information has the potential for being verified by former employers who are willing to answer straightforward job-related questions. A possible problem with this approach, however, is that it may discriminate against people who have not been given the opportunity to engage in certain behaviors in the past. Thus, checks for adverse impact would need to be conducted. When adverse impact is not a problem, it would appear likely that the two methods would significantly increase the validity of the interview as a selection device.

A possible limitation of the present research is that the validation of the situational

interview in two of the three studies was confounded with a training program to minimize rating errors. However, on the basis of the positive results obtained for hourly workers in Study 1, in which only a warning was given to avoid rating errors when evaluating job performance, it is doubtful that much of the variance in the validity coefficients is explainable by this training. For example, Levine and Butler (1952) and Wexley, Sanders, and Yukl (1973) found that a lecture or warning to minimize rating errors had little or no effect on rater behavior. It is likely that the need for extensive rater training was minimized in the present studies by the similarity of the interviewee answers with the benchmark answers. Nevertheless, Pursell, Dossett, and Latham (1980) have found that training raters to minimize rating errors when making performance appraisals can increase the validity coefficients of predictors significantly.

Reference Notes

1. Pursell, E. D. *Behavioral criteria for evaluating sawmill workers.* New Bern, N.C.: Weyerhaeuser Company, 1977.
2. Cohen, B. M., Moses, J. L., & Byham, W. C. *The validity of assessment centers: A literature review* (Monograph II). Pittsburgh: Development Dimensions Press, 1974.

References

Flanagan, J. C. The critical incident technique. *Psychological Bulletin,* 1954, *51,* 327–358.

Ghiselli, E. E. The validity of a personnel interview. *Personnel Psychology,* 1966, *19,* 389–395.

Lancaster, H. Failing system: Job tests are dropped by many companies due to antibias drive. *Wall Street Journal,* September 3, 1975, pp. 1.

Latham, G. P., Fay, C., & Saari, L. M. The development of behavioral observation scales for appraising the performance of foremen. *Personnel Psychology,* 1979, *32,* 299–311.

Latham, G. P., & Wexley, K. N. Behavioral observation scales for performance appraisal purposes. *Personnel Psychology,* 1977, *30,* 255–268.

Latham, G. P., Wexley, K. N., & Pursell, E. D. Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology,* 1975, *60,* 550–555.

Latham, G. P., & Yukl, G. A. A review of research on the application of goal setting in organizations. *Academy of Management Journal,* 1975, *18,* 824–845.

Levine, J., & Butler, J. Lecture versus group discussion in changing behavior. *Journal of Applied Psychology,* 1952, *36,* 29–33.

Locke, E. A. Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance,* 1968, *3,* 157–189.

Maas, J. B. Patterned expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology,* 1965, *49,* 431–433.

Mayfield, E. C. The selection interview—re-evaluation of published research. *Personnel Psychology,* 1964, *17,* 239–260.

Pursell, E. D., Dossett, D. L., & Latham, G. P. Obtaining valid predictors by minimizing rating errors in the criterion. *Personnel Psychology,* 1980, *33,* 91–96.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology,* 1963, *47,* 149–155.

Thorndike, R. L. *Personnel selection.* New York: Wiley, 1949.

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

Wagner, R. The employment interview: A critical review. *Personnel Psychology,* 1949, *2,* 17–46.

Wernimont, P. F., & Campbell, J. P. Signs, samples, and criteria. *Journal of Applied Psychology,* 1968, *52,* 372–376.

Wexley, K. N., Sanders, R. D., & Yukl, G. A. Training interviewers in employment interviews. *Journal of Applied Psychology,* 1973, *57,* 233–236.

Received October 1, 1979 ■