The Critical Incident Technique Revisited

Michael G. Aamodt and Cheryl Reardon
Radford University
Wilson W. Kimbrough
University of Arkansas

The paper reviews the literature on the use of the critical incident technique (Flanagan, 1954). The review demonstrates that the critical incident technique (CIT) is useful for a number of purposes; those include job analysis, training, and performance appraisal. The paper also concludes that the CIT is rather robust, since the methods used to generate and sort the incidents do not greatly affect the outcome of the finished product. However, variables such as tests for criticality, tests for category acceptance, and individual differences of incident generators do affect the outcome of the CIT procedure.

Introduction

The critical incident technique (CIT) was formally developed and reported by John Flanagan (1954) at the University of Pittsburgh. This technique prescribes a systematic set of procedures for collecting direct observations of human behavior (called critical incidents) which have made the difference between successful and unsuccessful job performance (Flanagan, 1954). Observations of critical incidents can be obtained in many ways, depending on the intended use of the observations. However, the usual method is to have job incumbents report, either through log books or through recall, at least one example of either excellent or poor job performance. These incidents are then sorted into categories which reveal the important dimensions within a job.

Critical incidents have become an important tool in the management of human resources in the field of police and criminal psychology since they are the basis for a major method of job analysis and the basis for at least five major methods of performance appraisal: Behaviorally Anchored Rating Scales, Behavioral Observation Scales, Behavioral Expectation Scales, Mixed Standard Scales, and forced choice rating scales. In addition, critical incidents by themselves are used as another major method of performance appraisal in which supervisors are asked to periodically record examples of job

performance for each employee. Finally, critical incidents have been suggested as a means toward the creation of a structured employment interview (Latham, Saari, Purcell, & Campion, 1980) and as a method of training (Glickman & Vallance, 1958; O'Brien & Plooij, 1977).

Even though the use of critical incidents has become a common and important tool in the human resource professional's armamentarium, relatively little research has been conducted on the optimal ways in which to collect and then sort these incidents. While there is a basic procedure for using critical incidents for job analysis and performance appraisal purposes, the specific procedures seem to be idiosyncratic to the person who is supervising the incident collection.

There are many examples for this lack of a standardized procedure in the use of the CIT. A review of the CIT literature indicates that the number of people used to generate the incidents has ranged from 11 to 3,767; the number of individuals used to sort the incidents has ranged from 1 to 110; five studies had the same individuals both generate and sort the incidents; and 23 studies used individuals from more than one job level to generate the incidents.

While Flanagan (1954) indicated that there was no standard method in which the CIT must be conducted, it appears that some methods are more adequate than others. Perhaps a reason for the lack of standardization in procedures is that little recent research has been conducted as to the optimal techniques used to collect and then sort the incidents. Therefore, it is the purpose of this paper to review and synthesize the available research on the CIT.

General Critical Incident Studies

The following pages review the studies which investigate the techniques which are involved in conducting the CIT. These studies have focused on many factors, including: the procedure by which the incidents are collected, the person generating the incidents, the number of incidents that are needed, the sorting procedure, and number of sorters needed.

The Method in Which Incidents Are Collected

Research in this area has focused on two variables, the time period in which the incidents are collected and the method used to collect the incidents. The only study that has investigated the first variable was by Nagay (1949), who studied air traffic controllers and found that the types of incidents generated by incumbents were affected by the season. That is, some job behaviors were important in the winter, but not in the summer. This finding would suggest that the researcher be cautious about collecting incidents during a small time period.

In addition to this, there appeared to be selective recall of dramatic or special types of incidents. This selective recall occurred when the incidents were reported several months after they had happened. Thus, it was recommended that log books be used to collect the incidents that would be collected over a period of time. This recommendation was also supported by Miller and Flanagan (1950), who discovered that foremen reporting incidents weekly forgot twice as many incidents as did foremen reporting incidents daily. However, neither Flanagan, Miller, Burns, Hendrix, Stewart, Preston, and West (1953) nor Campion, Greener, and Wernli (1973) found any differences in incidents collected through an interview procedure when compared to those collected using log books. Finkle (1950) found that incidents obtained through questionnaire booklets yielded similar results to incidents obtained through individual interviews. Wagner (1950) extended these findings to indicate that

incidents obtained through group interviews were comparable to those obtained through individual interviews. Finally, Wagner (1951) found that incidents collected by one interviewer were similar to those obtained by other interviewers. Thus, it would appear that the method used to collect the incidents: log book, group interview, or individual interview, does not have a major impact on the outcome of the CIT.

Wording of the Incident Request

Even though Flanagan (1954) stated that slight changes in the wording of the incident request will lead to changes in the types of incidents that will result, the authors of this review were unable to find research that would support this contention. In the two studies that investigated this issue, Finkle (1950) and Mullins (1983) asked subjects to report incidents that represented either a slight deviation or a substantial deviation from normal behavior. Both studies revealed only a slight difference in the types of incidents provided in each condition.

Latham and Marshall (1982) looked at the effect of goal setting on the number of incidents generated by government employees and found that including a goal in the instructions increased the number of incidents generated. In this study, employees were placed into one of three goal setting conditions: self-set goals, participatively set goals, or assigned goals. The results indicated that none of the goal setting methods was superior to another in terms of the number of critical incidents that were produced by each employee. However, the higher the goal that was set, the greater the number of incidents generated.

Characteristics of the Person Generating the Incidents

The first study investigating this area was by Wagner (1950) who studied the critical requirements for dentists. Incidents were collected from patients, dentists, and dental school instructors and placed into one of four job categories. An analysis of the source by category frequency distribution revealed that the source of the incident affected the category in which the incident was placed. That is, patients reported more patient-dentist relationship incidents while dentists and instructors reported more technical proficiency and acceptance of professional responsibility incidents.

Compatible results were found by Smit (1952) who investigated differences in the types of incidents reported by psychology students and faculty; by Ronan and Latham (1974) who found differences between dealers and foresters; and by Andersson and Nilsson (1964) who collected 1,800 incidents from superiors, assistants, store managers, and customers in a Swedish grocery company. However, Weislogel (1952) found no differences in the patterns of critical requirements supplied by managers and agency heads at a life insurance firm.

The most recent studies concerning differences among the individuals generating the incidents were by Aamodt, Kimbrough, Keller, and Crawford (1982) and Aamodt (1983). Aamodt et al. (1982) investigated the relationship between the race, sex, and job performance level of individuals generating critical incidents and the types of incidents that each person generated. The results of this study indicated that the race of the person affected the frequency of incidents produced in the respective categories. Black Resident Assistants generated more incidents that were sorted into the two categories of "Interest in Residents" and "Fairness." Even though sex and job performance level were not related to incident generation, a number of problems with the sample may have masked any real differences.

In order to overcome some of the problems with the sample in Aamodt et al. (1982), (e.g. a restriction in range in the measure of job performance level), Aamodt (1983) investigated the relationships between the sex, job performance level (G.P.A.), and personality of general psychology students and the types of critical incidents that the students generated regarding effective and ineffective teaching. The results indicated that all three variables were related to incident generation. However, even though the relationships were significant, the effect sizes were fairly small.

Similar results were found by Machungwa and Schmitt (1983) when they investigated the use of the CIT as means toward understanding work motivation in a developing country. They found significant but small relationships between individual difference variables and the generation of critical incidents. Thus, the results of Aamodt et al. (1982), Aamodt (1983), and Machungwa and Schmitt (1983) indicate that personal variables slightly moderate the generation of critical incidents.

Number of Incidents Needed

Flanagan (1954) has suggested that 1,000-2,000 incidents are necessary for semi-skilled jobs and 2,000-4,000 incidents are needed for supervisory level jobs. However, the results of the Andersson and Nilsson (1964) study indicated that 95 percent of the categories appeared after sorting only two-thirds of the incidents. Thus, not all 1,800 incidents were needed. Similar to this, Jensen (1951) found that seldom would new types of behavior appear after reviewing 400 of 500 incidents.

The issue of the number of incidents needed was extended by Mullins (1983) to include the number of locations and generators needed. Mullins had 97 campus police officers at 13 universities generate critical incidents. The results indicated that no new incidents appeared after examining the incidents from the first three universities. Furthermore, after examining the incidents supplied by the first 19 incumbents, neither new incidents nor any new categories appeared.

Thus, it appears the Flanagan's suggestion, as to the number of incidents necessary, is not supported and that fewer incidents can be used. While fewer incidents may be needed, there is still no indication as to the minimal number of incidents necessary for full coverage of a job, although the number does appear to be much lower than 1,000.

Reliability of Incident Sorting

Andersson and Nilsson (1964) investigated the reliability and validity of the CIT and prefaced their article with the statement "Although the method has been used in a practical manner in hundreds of job analyses, relatively little has been done to study the method itself with respect to either reliability or validity" (p. 398).

Therefore, Andersson and Nilsson (1964) investigated the reliability issues by studying the extent to which individuals could easily place incidents into a category system. To do this, twenty-four students worked in pairs in order to "place a group of 100 incidents in corresponding categories" (p. 401). The results indicated that the pairs of incumbents agreed 80 percent of the time about which of three categories an incident should be placed into. This agreement rate decreased to around 66 percent when the number of categories was increased to 17. Similar sorting agreement results were also found by Bridgman, Spaeth, Driscoll, and Fanning (1958), Aamodt et al. (1982), Ronan and Latham (1974), and Machungwa and Schmitt (1983). However, these results were not upheld by either Lowenberg (1979) or Aamodt (1983).

It should be pointed out that both Aamodt (1983) and Lowenberg (1979) also investigated the possibility that group differences may occur in sorting critical incidents. That is, do males sort incidents into different categories than do females? The results in both studies seem to indicate that no such differences exist. In addition to this, Blankenship and O'Brien (1983) found that the gender of the person described in the incident did not affect the relative placement of incidents into the respective categories.

Number of Sorters That Are Needed

The only study looking at this variable was Aamodt (1983) who compared groups of 3, 6, 12, 24, 48, and 100 sorters and found that the relative numbers of incidents per category were similar regardless of the number of sorters that were used. Thus, for economic reasons, a researcher might be justified in using as few as three sorters.

Sorter Agreement Level

Various researchers have used sorter agreement levels ranging from just over 50 percent all the way to 100 percent. However, results from Aamodt (1983) and Bernardin, LaShells, Smith, and Alvares (1976) indicate, respectively, that agreement levels of 62 percent yield results similar to agreement levels of 75 percent and that agreement levels of 60 percent yield results similar to agreement levels of 80 percent. Thus, a researcher can be confident that use of agreement levels of around 60 percent will be comparable to higher agreement levels.

The Test for Criticality

It has been suggested that before sorting, incidents should be examined to determine if the effective incidents are actually examples of effective performance, and if the ineffective incidents are actually examples of ineffective performance. If an incident does not pass this test for criticality, it should be discarded. Only two studies found by the authors have used such a test. The first study by Lowenberg (1979) applied the test of criticality to her incidents and found that 16 percent of the incidents had to be discarded because their direction of criticality was ambiguous. The second study, by Van Fleet (1974), found that 47 percent of his incidents did not pass this criticality test. Therefore, when it is important to distinguish between the number of incidents that are examples of effective and ineffective behavior, it seems reasonable to apply this test of criticality. An example of such an instance would be when incidents are used for training or performance appraisal.

It is at this point in the review that it might be a good idea to briefly discuss the use of effective versus ineffective incidents. In many studies reviewed by the authors, incidents of effective and ineffective behavior were combined into one category. However, on the basis of his data and the data of others, Aamodt (1983) had suggested that the relative numbers of incidents in the respective categories are different for ineffective and effective incidents. While it is not yet known what effect this difference might have on the outcome of a job analysis or training manual, the fact that those numbers are often different suggests that until further research is done, effective and ineffective incidents should be considered separately.

The above review of the CIT literature has been limited to studies that were concerned with the generation and sorting of critical incidents regardless of their intended use. Because critical incidents are used for many purposes, a separate discussion of the main functional areas for critical incidents is necessary. Thus, a discussion of each of the major critical incident uses follows.

Major Critical Incident Uses

Job Analysis

When Flanagan (1954) published his classic article on the CIT, the primary purpose of the technique was job analysis. However, enthusiasm for the use of the CIT as a job analysis tool seems to have waned in recent years. Levine, Bennett, and Ash (1979) surveyed 106 public personnel practitioners, to determine the extent to which the various major job analysis techniques were used. The results indicated that the CIT was only the third most popular method, trailing both *task analysis* and *job elements*. It should be kept in mind that this study examined only the major structured job analysis methods. Jones and DeCotiis (1969) surveyed companies and discovered that the most common type of job analyses used was the interview and the job observation. The CIT did not even rank.

It is difficult to determine the reasons for the differences found between the two studies. One reason could be that the form of the two questionnaires was different, which led to different types of responses. A second reason could be that the samples were not equivalent. A final reason centers around the temporal difference between the two studies. It could be that the use of critical incidents had actually increased in the ten years separating the two studies. Regardless of the relative use of the CIT in job analysis, the absolute rate appears to be rather low.

One possible reason for this low rate of use might be found in the Levine et al. (1979) study. When subjects were asked about the formal training that they had received in job analysis, the replies were as follows: 44 percent had received training in job elements; 42 percent had received training in task analysis; nine percent had recieved training in the CIT; and seven percent had received formal training in the use of the Position Analysis Questionnaire (PAQ).

Another explanation for the low use of the CIT is that personnel practitioners evaluated task analysis and job elements as potentially being the most effective, and evaluated both critical incidents and PAQ as potentially being the least effective. This finding could be explained by a cognitive dissonance explanation—that individuals who use a particular method would be more inclined to rate it as being effective in order to cognitively justify its use.

In a follow-up study that seems to be the most revealing study in its area, Levine, Ash, and Bennett (1980) had 64 personnel professionals analyze four job classes using four job analysis mehtods (job elements, task analysis, PAQ, and CIT). The results of the study indicated a number of interesting findings. First, the length of the job analysis reports varied to a great degree. Job elements reports took an average of 12 pages to complete, the PAQ report and the CIT reports took and average of 20 pages to complete, and the task analysis reports took and average of 27 pages to complete.

Second, the CIT was rated most adequate for producing performance measures information. Job elements was rated both the most appropriate for measuring specific job components and the most useful for content validity. The PAQ was rated overall as the least effective method but was also considered the easiest to use. Finally, the CIT and task analysis were the most expensive methods to use.

While the previously mentioned results appear favorable to the CIT, Levine, Ash, Hall, and Sistrunk (1983) have recently reported less favorable results. Levine et al. described seven job analysis techniques to 93 job analysis experts and asked them to rate each method's effectiveness. The CIT received low ratings on all seven variables.

However, it should be noted that the study only *described* job analysis techniques and used survey methodology. The results might have been different, had the ratings been of finished job analysis projects.

In a study investigating the validity of the CIT as a job analysis tool, Ronan and Latham (1974) collected data on wood producers' production, turnover, absenteeism, and injuries. They then compared these critiera with the types of critical incidents that were written about each producer. The results indicated that "it is possible to predict the job performance of producers on various criteria on the basis of observations of behaviors derived from critical incidents" (p. 61). Thus, Ronan and Latham (1974) found support for the CIT as a reliable and valid means of job analysis.

Training

Even though critical incidents would intuitively seem to have potential as a training tool, little research has explored this avenue. The major investigation into the use of critical incidents for training was conducted by Glickman and Vallance (1958) who explored the CIT as a method to assess naval officer needs. Glickman and Vallance collected 1,073 incidents of both superior and poor performance by naval ensigns. The incidents were then sorted by officer-instructors into eight categories which paralleled the current naval training curriculum. Within each category were two subcategories, "taught" and "not taught." The sorter would first read each incident and then decide into which category the incident belonged. If it was behavior currently covered in training, the incident was put into the "taught" subcategory. If it was behavior that was not currently being covered, it was put into the "not taught" subcategory.

The results indicated that 63% of the incidents were sorted into the "taught" subcategory, indicating that even though much of the critical job behavior was covered in training, there was still room for improvement.

The next procedure used by Glickman and Vallance (1958) was to ask job experts to estimate how soon after reporting aboard the ship was a new officer expected to handle either sort of situation or demonstrate behavior described in the incidents. The results indicated that the ineffective behavior in the incidents was corrected in a significantly shorter time after reporting than the effective behavior that was described in the incidents was learned. Furthermore, the investigators found a significant positive correlation (r=.65) between the number of incidents per category and the estimated time needed for satisfactory performance. The authors concluded that the ineffective behaviors need to be corrected before the effective behaviors are learned, and that the number of incidents that are sorted into each category can be used as a means of estimating the importance of the categories.

The important finding in the article was that the critical incident procedure utilized by Glickman and Vallance (1958) could be used as a means for evaluating the content validity of a training program. The greater the percentage of incidents sorted into the "taught" subcategory, the greater the amount of confidence in the adequacy of a training program.

Another important finding is that it appears that the number of incidents sorted into each category can be used as a means for determining the importance of job dimensions. Recent support for this last finding was reported by Aamodt and Kimbrough (1985) when they had incumbents rank order the importance of job dimensions that were discovered in a job analysis. The results indicated a significant correlation between the number of incidents gener-

ated per category and the category importance ranking.

A second study involving critical incidents and training need assessment was conducted by Folley (1969). Folley collected over 2,000 incidents of behavior by department store sales personnel and suggested that categories receiving the most number of incidents be given extra emphasis during training.

Perhaps the most creative use of critical incidents for training was demonstrated in a study by O'Brien and Plooij (1977) who used critical incidents to create a training manual for nurses working with Aboriginals in South Australia. The authors had 60 workers describe incidents of effective and ineffective performance. "Major incidents were selected on the basis of the frequency of their occurrence and the degree of agreement in their interpretation" the authors pointed out (p. 500). This procedure resulted in 41 incidents being placed into a manual with discussions on each incident situation. Subjects receiving this manual were compared to subjects who had received either a manual containing essays written by experts or who had received no manual at all. The results of the study indicated that subjects who used the critical incident manual had greater retention and generalization of cultural knowledge than subjects in the other groups.

Employee Selection

One interesting use of the CIT stems from an article by Latham, Saari, Purcell, and Campion (1980) when they used critical incidents to form the basis for a situational interview. Latham et al. (1980) first conducted a critical incident job analysis of sawmill workers and foremen. These incidents were rewritten for clarity and suffixed with the question "What would you do in this situation?" In the course of the employment interview, applicants were read the various situations and asked what they would do. The answers to these situational questions were recorded and then rated by supervisors on the five point scale using previously given benchmark answers as a guide.

The results of the study indicated that interrater reliabilities were in the high .70s. In addition, interview scores produced concurrent validity coefficients ranging from .28 to .51 and predictive validity coefficients in the .30s. The authors concluded that the situational interview was a promising method of employee selection.

Performance Appraisal

In the field of performance appraisal there are two major uses of the CIT. The first of these methods centers around the supervisor recording examples of the critical behavior that he/she observes in each employee. Then, after a period of time (usually once every six months), the supervisor has a conference with the employee to discuss the employee's job performance. The critical incidents are then used to support the subjective ratings that the supervisor communicates to the employee. While many texts (e.g. McCormick & Ilgen, 1980), discuss this as a popular method of performance appraisal, studies investigating this method are difficult to find.

The two most often cited articles that reviewed the critical incident method of performance appraisal were written by Oberg (1972) and by Flanagan and Burns (1955). The article by Oberg (1972) mentioned that the two biggest advantages to the CIT were that, a) during the performance review interview, the discussion revolves around actual behavior rather than worker traits, and b) that the employee's performance rather than his/her personality was what was being criticized or praised. The two biggest disadvantages given were that the process of gathering the incidents can become a chore and that the use of the critical incident method may cause a

supervisor to delay feedback to an employee.

In spite of these problems, Flanagan and Burns (1955) collected over 2,500 critical incidents and through a sorting procedure, reduced the incidents down to 16 critical job requirements for hourly wage employees. These 16 requirements were used to create an "employee performance record" that the supervisor could use to keep track of the critical behavior for his/her employees. The important findings that were noted by Flanagan and Burns were that daily recording of incidents led to twice as many incidents being recorded as compared to weekly recording and five times as many incidents when compared to bi-monthly recording. The recording of daily incidents usually took less than five minutes to do and showed that over 90% of the incidents recorded were positive and that about 25% of the employees had no incidents recorded about them. Finally, it was found that after four years of recording incidents, employee suggestions had increased by 100% and disciplinary warnings were reduced by half. Unfortunately, there was no control group with which these numbers could be compared.

The second major use of critical incidents in performance appraisal involves Behaviorally Anchored Rating Scales (BARS). The concept of BARS was originally introduced by Smith and Kendall (1963) as a means of reducing common rating errors by anchoring each point on a graphic rating scale with an example of job behavior. While the actual utility of BARS is questionable (Kingstrom & Bass, 1981), their popularity (at least in terms of research) cannot be disputed. Although there are various ways in which BARS are constructed, the basic steps are as follows:

1. Obtain critical incidents.
2. Examine and cluster the incidents into categories.
3. Sort each incident into the categories found in step two.
4. Rate each incident as to its level of job importance.
5. Retain the incidents whose ratings had low standard deviations and whose scale value approximates one of the points on the graphic rating scale.

A review of the BARS literature indicates that little research has investigated the optimal developmental procedures to use in the collection and retranslation of critical incidents. For example, there has been little consistency and much variation in the way and the number of incidents generated. Burnaska and Hollman (1979) used 33 students and had each generate three examples of both effective and ineffective behavior. Motowidlo and Borman (1977) obtained 1,163 incidents from 190 generators, and Shapira and Shirom (1980) were able to obtain 222 incidents from 37 generators.

There is also a great variation in the number of incidents generated, as well as in the type of persons generating the incidents. In a survey of 32 BARS articles, the number of individuals generating the incidents ranged from 8 to 376 with a median of only 37. Nine studies set standards for the number of incidents that were to be generated by each person, while the other studies set no standards. In these later studies, the average number of incidents generated per person ranged from 6.0 to 15.2. Most studies did not state the number of individuals generating the incidents, the number of incidents generated, the number of sorters used, the sorting agreement level, and so forth. In other words, it is not certain as to how the scales were created. This lack of information is especially important when an article shows no advantage of BARS when compared to other scales. It is possible that poor critical incident

methodology is the culprit, and not the type of scale.

On the brighter side, BARS research has led to the contribution of three novel findings in the use of the CIT. The first of these findings was by Bernardin, LaShells, Smith, and Alvares (1976) when they investigated the effect of using a 60% sorter agreement level versus an 80% sorter agreement level. The results indicated that BARS constructed with a 60% agreement level were no less reliable than BARS constructed with an 80% agreement level.

The second finding, also by Bernardin et al. (1976), was that BARS were more reliable when one group retranslates the incidents and another group scales the incidents. Thus, the same people should not be used to generate, sort, and scale the critical incidents.

The third important contribution to critical incident knowledge was provided by Campion, Greener, and Wernli (1973) who had one group of incumbents recall behavioral examples and has another group use log books to record incidents as they occurred. The authors concluded that recall and work observation were essentially equivalent both in terms of their reliability and their generalizability across rater groups.

Summary and Conclusion

As revealed in the previous pages, the CIT can be a valuable tool with many potential uses. However, research on the CIT has barely scratched the proverbial "tip of the iceberg." It seems important that research efforts be intensified in order to first discover the optimal ways in which the technique should be conducted (if, indeed, there are optimal ways) and then to compare the utility of the CIT to the utilities of more popular methods.

Until further research reveals new findings, the following conclusions can be made concerning the use of the CIT:

1. The method used to collect the incidents (log book, interview, questionnaire) does not appear to affect the outcome of the CIT.
2. The wording of the incident request does not appear to affect the outcome of the CIT.
3. A representative sample of all types of people involved in the job (incumbents, supervisors, etc.) is essential.
4. Less than 500 incidents are probably needed to obtain full coverage of a job.
5. Incidents can be sorted reliably.
6. Three sorters appear to be all that are needed.
7. Tests for incident criticality and category acceptance should be used.
8. Incidents of ineffective and effective behavior should be analyzed separately.
9. Critical incidents can be used for many different purposes ranging from job analysis to training.

Referenc

Aamodt, M.G. (1983). *Relationship between sex, personali* *rformance level and the generation*
and sorting of critical incidents. Unpublished doctoral d University of Arkansas, Fayetteville.

Aamodt, M.G., & Kimbrough, W.W. (1985). Comparison of ds for weighting multiple predictors.
Educational and Psychological Measurement, 45, 477–

Aamodt, M.G., Kimbrough, W.W., Keller, R.J., & Craw (1982). Relationship between sex,
race, and job performance level and the generation of cr ds. *Educational and Psychological*
Research, 2, 227-234.

Andersson, B.E., & Nilsson, S.G. (1964). Studies in the nd validity of the critical incident
technique. *Journal of Applied Psychology, 48,* 398–403.

Bernardin, J.J., LaShells, M.B., Smith, P.C., & Alvares, 6). Behavioral expectation scales:
Effects of developmental procedures and formats. *Journ* *d Psychology, 61,* 75-79.

Blankenship, D., & O'Brien, B. (1983). *Effect of gender o* *n critical incidents on the sorting*
of critical incidents. Paper presented at the Fifth Ann B. Graduate Student Conference,
Virginia Beach, Virginia.

Bridgman, C.S., Spaeth, J., Driscoll, P., & Fanning, J. (1 men helped by bringing out jobs'
critical incidents. *Personnel Journal, 36,* 411-414.

Burnaska, R.F., & Holmann, T.D. (1979). An empirical con he relative effects of rater response
biases on three rating scale formats. *Journal of Applied* *59,* 307-312.

Campion, J.E., Greener, J., & Wernli, S. (1973). Work obse as recall in developing behavioral
examples for ratings scales. *Journal of Applied Psycholo* 288.

Finkle, R.B. (1950). A study of the critical requirements of for *University of Pittsburgh Bulletin,*
46, 291-297.

Flanagan, J.C. (1954). The critical incident technique. *Psych* *etin, 51,* 327-358.

Flanagan, J.C., & Burns, R.K. (1955). The employee performa A new appraisal and development
tool. *Harvard Business Review, 33,* 95-102.

Flanagan, J.C., Miller, R.B. Burns, R.K., Hendrix, A.A., S , Preston, H.O., & West, E.D.
(1953). *The employee performance record for non-super* *ied employees.* Chicago: Science
Research Associates.

Folley, J.D. (1969). Determining training needs of departn sales personnel. *Training and*
Development Journal, 23, 24-26.

Glickman, A.S., & Vallance, T.R. (1958). Curriculum assessmen cal incidents. *Journal of Applied*
Psychology, 42, 329-335.

Jensen, A.C. (1951). Determining critical requirements for teac *al of Experimental Education,*
20, 79-86.

Jones, J.J., & DeCotiis, T.A. (1969). Job Analysis: National ings. *Personnel Journal,* 805-
890.

Kingstrom, P.O., & Bass, A.R. (1981). A critical analysis of mparing Behaviorally Anchored
Rating Scales (BARS) and other ratings formats. *Personnel* *, 34,* 263-289.

Latham, G.P., & Marshall, H.A. (1982). The effects of self- atively set and assigned goals
on the performance of government employees. *Personnel Psy* , 399-404.

Latham, G.P., Saari, L.M., Purcell, E.D., & Campion, M.A. (situational interview. *Journal*
of Applied Psychology, 65, 422-427.

Levine, E.L., Ash, R.A., & Bennett, N. (1980). Exploratory comp y of four job analysis methods.
Journal of Applied Psychology, 65, 524-535.

Levine, E.L., Ash, R.A., Hall, H., & Sistrunk, F. (1983). Evaluati alysis methods by experienced
job analysis. *Academy of Management Journal, 26,* 339-348.

Levine, E.L., Bennett, N., & Ash, R.A. (1979). Evaluation an our job analysis methods for
personnel selection. *Public Personnel Management, 8,* 146-151.

Lowenberg, G. (1979). Interindividual consistencies in determinin -based dimensions of teaching
effectiveness. *Journal of Applied Psychology, 64,* 492-501.

Machungwa, P.D., & Schmitt, N. (1983). Work motivation in a developing country. *Journal of Applied Psychology, 68,* 31-42.

McCormick, E.J., & Ilgen, D.R. (1980). *Industrial psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Miller, R.B., & Flanagan, J.C. (1950). The performance record: An objective merit-rating procedure for industry. *American Psychologist, 5,* 331-332.

Motowidlo, S.J., & Borman, W.C. (1977). Behaviorally anchored scales for measuring morale in military units. *Journal of Applied Psychology, 62,* 177-183.

Mullins, W.C. (1983). *Job analysis outcomes as a function of group composition.* Unpublished doctoral dissertation, University of Arkansas, Fayetteville.

Nagay, J.A. (1949). *The development of a procedure for evaluating the proficiency of air route traffic controller.* Washington, D.C.: Civil Aeronautics Administration.

Oberg, W. (1972). Make performance appraisal relevant. *Harvard Business Review, 50,* 61-67.

O'Brien, B.E. & Plooij, D. (1977). Comparison of programmed and prose culture training upon attitudes and knowledge. *Journal of Applied Psychology, 62,* 499-505.

Ronan, W.W., & Latham, G.P. (1974). The reliability and validity of the critical incident technique: A closer look. *Studies in Personnel Psychology, 6,* 53-64.

Shapira, Z., & Shirom, A. (1980). New issues in the use of Behaviorally Anchored Rating Scales: Levels of analysis, the effects of incident frequency, and external validation. *Journal of Applied Psychology, 65,* 517-523.

Smit, J.A. (1952). A study of the critical requirements for instructors of general psychology courses. *University of Pittsburgh Bulletin, 48,* 279-284.

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

Van Fleet, D.D. (1974). Toward identifying critical elements in a behavioral description of leadership. *Public Personnel Management, 3,* 70-82.

Wagner, R.F. (1950). A study of the critical requirements for dentists. *University of Pittsburgh Bulletin, 46,* 331-339.

Wagner, R.F. (1951). Using critical incidents to determine selection test weights. *Personnel Psychology, 4,* 373-381.

Weislogel, M.H. (1952). Critical requirements for life insurance agency heads. *University of Pittsburgh Bulletin, 48,* 300-305.