McMASTER
• U N I V E R S I T Y •

MICHAEL G. DeGROOTE
SCHOOL OF BUSINESS

RESEARCH AND
WORKING PAPER
SERIES

THE CONTRIBUTIONS OF JOB RELEVANCE, TIMING, AND RATING SCALE TO THE VALIDITY OF THE STRUCTURED EMPLOYMENT INTERVIEW

By

Willi H. Wiesner
Michael G. DeGroote School of Business
McMaster University
Hamilton, Ontario, Canada

Working Paper # 381

September, 1992

ER UNIVERSITY
Street West
)ntario, Canada L8S 4M4
416) 525-9140

THE CONTRIBUTIONS OF JOB RELEVANCE, TIMING, AND RATING SCALE TO THE VALIDITY OF THE STRUCTURED EMPLOYMENT INTERVIEW

By

Willi H. Wiesner
Michael G. DeGroote School of Business
McMaster University
Hamilton, Ontario, Canada

Working Paper # 381

September, 1992

The Contributions of Job Relevance, Timing, and Rating Scale to the Validity of the Structured Employment Interview

Willi H. Wiesner

Michael G. Degroote School of Business

McMaster University

Hamilton, Ontario

This paper is based on the author's Doctoral Dissertation research which was supported by a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada.

Abstract

The investigation addressed the problem of improving the criterion-related validity and reliability of the selection interview in the context of a controlled experiment, using unconventional methodology. Interview ratings made by subjects viewing videotaped employment interviews were validated against job performance ratings provided by the interviewees' supervisors. The job relevance of interview questions, the timing of interview ratings, and the type of rating scale used to rate interviewees' answers were found to have a significant impact on the reliability and validity of the interview.

The employment interview is one of the oldest and most widely used of all selection procedures. However, although the interview has remained popular among practitioners, until recently reviews of employment interview literature have been pessimistic about the employment interview as a reliable or valid selection device (Arvey & Campion, 1982; Mayfield, 1964; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949; Webster, 1982; Wright, 1969). Nevertheless, employers have refused to abandon the interview and, indeed, Arvey and Campion (1982) point out that some employers have abandoned selection tests in favour of the interview when faced with the prospect of equal rights litigation. Such employers have failed to recognize that the interview is generally an even less defensible selection device than most tests (Arvey, 1979; Arvey & Campion 1982; Rowe, 1981). In light of employers' tenacious adherence to the interview in the face of possible litigation, research on the employment interview has continued to be vitally important.

Although reviewers have generally been unimpressed by the criterion-related validity of the employment interview, they have recently suggested that certain kinds of interviews (i.e., structured interviews) have greater validity than others (i.e., unstructured interviews) (Arvey & Campion, 1982; Harris, 1989; Webster, 1982). In fact, two recent meta-analyses of employment interview literature support this observation (McDaniel, Whetzel, Schmidt, Hunter, Maurer, & Russel 1987; Wiesner & Cronshaw, 1988). Wiesner and Cronshaw (1988) found that, although employment interviews had acceptable validity overall (average uncorrected validity coefficient of .26 across the 150 effect sizes they analyzed) structured interviews have significantly higher validity (average uncorrected coefficient of .34) than unstructured interviews (average uncorrected coefficient of .17). Moreover, although all of the variance in validity coefficients of unstructured interviews was accounted for by the effects of statistical artifacts, much of the variance in validity coefficients of structured interviews remains unexplained, suggesting the operation of one or more moderators. Simply put, the structured interviews used in some studies were better predictors of job performance than in others, begging the question of what factors might contribute to these differences.

An examination of interview literature reveals considerable differences in what is considered to be a "structured" interview. Initial versions of structured interviews were simply standardized interviews. That is, the same questions were asked of all interviewees for a particular position. These questions were to be read from a list which had been prepared in advance but minor deviations from or additions to the list were permitted (Bass, 1951; Bingham & Moore, 1931; Fear, 1958; Fear & Jordan, 1943; Hovland & Wonderlic, 1939; Kenagy & Yoakum, 1925; McMurry, 1947; Rundquist, 1947; Viteles, 1932; Yonge, 1956). More recent approaches to structured interviewing such as the Situational Interview (Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980) and the Behavior Description Interview (Janz, 1982) retain the principle of standardizing questions, although they go considerably beyond it. The standardization of interview questions might be viewed, then, as the minimal condition for an interview to be considered "structured" (Webster, 1982).

Given the broad definition of a structured interview as one that is at least standardized, there are still considerable differences in the way various researchers structure their interviews. Whereas some researchers rely on a formal job analysis for the construction of the interview (e.g., Janz, 1982; Latham et al., 1980), a formal job analysis is not necessarily used for all interviews considered to be "structured" (e.g., Walsh, 1975). Even the Janz (1982) and Latham et al. (1980) approaches differ in a number of important ways. The Behavior Description Interview, for example, focuses on past behaviors and has used note-taking as a means of collecting information obtained from the applicant. The Situational Interview, on the other hand, focuses on intended behaviors in hypothetical situations and uses rating scales with scoring guides. Such differences in the way interviews are structured are likely responsible, at least to some degree, for the variability in the validity coefficients of structured interviews.

The differences in interview "structuring" techniques that have been used by various researchers suggest three potential moderators of interview validity (among others): the job relevance of interview questions, the timing of interview ratings, and the kind of scoring process or rating scale used. The purpose of this study is to explore the roles of these three variables in enhancing interview reliability and validity. A brief rationale for considering these variables is presented below.

Firstly, interviews often contain questions of little job relevance which provide interviewers with the kind of information that can feed pre-existing biases and stereotypes without revealing much about the applicant's ability to do the job (Webster, 1982). Eliminating such questions and focusing questions only on issues of direct relevance to the job should reduce the degree to which irrelevant information plays a role in the selection decision and increase the relative impact of relevant information. That is, as interview questions become more job-relevant, the impact of irrelevant information on the interview decision is reduced and the interview should have greater reliability and validity (Heneman, Schwab, Huett, & Ford, 1975; Landy, 1976; Langdale & Weitz, 1973; Osburn, Timmreck, & Bigby, 1981; Wiener & Schneiderman, 1974). Job relevance can be optimized by basing the interview questions on a formal job analysis.

Secondly, although some proponents of structured interviewing advocate rating candidates at the end of the interview (e.g., Fear, 1978), a number of researchers suggest that interviewees' answers should be scored as soon as they are given (Carlson, Thayer, Mayfield, & Peterson, 1971; Maas, 1965; Mayfield, Brown, & Hamstra, 1980). Delay in scoring the answers may allow the scoring to be influenced by errors in the interviewer's retrieval of information from memory, particularly if the notes omit important detail. Moreover, information retrieved from memory is more likely to be affected by stereotyping and the biases the interviewer holds with respect to the interviewee than information coded and scored immediately it is provided (Arvey, 1979; Webster, 1982). Rating answers to interview questions as soon as they are given should therefore yield higher interview validity than waiting until after the interview.

Thirdly, some structured interviews involve the rating of dimensions such as appearance and work history or traits such as motivation to work rather than rating the answers to individual questions (e.g., Heneman, Schwab, Huett, & Ford, 1975; Landy, 1976). Even when individual answers are rated, they are often rated using graphic rating scales (e.g., a scale from 1 to 5 with 1 representing "poor" and 5 representing "excellent" or even simpler versions). A major disadvantage of the numerical or adjective rating scales is that raters using such scales often disagree on the meanings of different rating levels. One rater's rating of "4" might be equivalent to another rater's "2" or "3" on the same trait or dimension. Maas (1965) advocated the use of Smith and Kendall's (1963) behavior expectation scales, rather than numerical or adjective rating scales, to rate interviews. Behavior expectation scales consist of benchmark answers for each scale value. What is meant by a "1" or a "3" is defined for the interviewer in terms of differentially effective work behaviors, one of which the applicant would be expected to engage in if he or she were hired. Maas found significantly higher inter-rater reliability when the behavior expectation rating method was used (although his interviewers still rated traits rather than interviewees' answers to interview questions). Latham et al. (1980) have adapted behavior expectation scales in designing scoring guides for their Situational Interview. The applicant's answer is scored with reference to a scoring guide which is developed using the Critical Incidents Technique (Flanagan, 1954). Examples of job-related behaviors that varied in effectiveness in particular situations are collected and refined to serve as sample answers in the scoring guide. Thus, numerical values on the scale are illustrated with examples of answers that would be worth a "1" or a "3" or a "5". Latham et al. and others have obtained significant validity coefficients using this approach. Using rating scales with scoring guides or benchmark answers should result in higher inter-rater reliability and, therefore, validity than using simple graphic rating scales to score answers to interview questions (Maas, 1965; Osburn et al., 1981).

Of the three potential moderators considered above, it seems reasonable that the job-relevance factor should have the greatest impact on the validity of the interview. If interview questions elicit information of little relevance, the manner in which this irrelevant information is scored should make little difference. The timing of interview scoring should follow in importance. If answers to interview questions cannot be remembered accurately, the kind of rating scale used to score the poorly remembered answers should have relatively little impact on validity. Interviewers rating answers at the end of the interview should have some difficulty retrieving relevant information from memory. Finally, the type of rating scale used should be the least important factor, although it should still have an effect on interview validity.

In summary, the effects of three potential moderators on structured interview validity are examined in this study: the job relevance of questions, the timing of ratings, and kind of rating scale used. The three variables are dichotomized (high vs. low job relevance of questions; rating answers during vs. after the interview; using a scoring guide vs. a simple graphic rating scale) to create a simplified 2 X 2 X 2 experimental design. However, there is no justifiable reason for including two of the possible conditions in the above design. The two conditions which are excluded from further consideration are the low job-relevance conditions in which a scoring guide would be used. These two conditions do not make any practical sense because it is not possible to devise a meaningful scoring guide for interview questions which have low job relevance (e.g., "What are your hobbies?", "What kind of people do you like best?", etc.). Thus, six experimental conditions remain in the design (see Table 1).

In light of the review provided above, the following hypotheses are proposed:

Hypothesis 1. Validity coefficients will be aligned from greatest to smallest across the six conditions as follows: (1) questions are highly job-relevant and answers are scored during the interview using a scoring guide; (2) questions are highly job-relevant and answers are scored during the interview using a graphic rating scale (i.e., numeric anchors 1 to 5); (3) questions are highly job-relevant and answers are scored after the interview using a scoring guide; (4) questions are highly job-relevant and answers are scored after the interview using a graphic rating scale; (5) questions are low in job-relevance and answers are scored during the interview using a graphic rating scale; (6) questions are low in job-relevance and answers are scored after the interview using a graphic rating scale.

Hypothesis 2. Inter-rater reliability coefficients will follow the same order of magnitude across conditions as that specified for validity coefficients in hypothesis 1 above. Reliability will be greatest when questions are highly job-relevant and answers are scored during the interview using a scoring guide and it will be least when questions have low job relevance and answers are scored after the interview using a graphic rating scale.

Although there are several different approaches to structured interviewing that could be used, the Behavior Description Interview (Janz, 1982; Janz, Hellervik, & Gilmore, 1986) was chosen as a vehicle to test the hypotheses because it is an approach with which the author was familiar. Moreover, no formal scoring guide has been used to rate interview answers in the Behavior Description Interview (BDI). This study will therefore examine how scoring guides might contribute to the reliability and validity of the BDI in particular, as well as possible implications for other approaches to structured interviewing.

Method

The present study is essentially a validation study in which interview ratings are evaluated as predictors of job performance. However, unlike the traditional validation study in which a large number of interviewees are assessed by a few interviewers, in this investigation interviews with a few interviewees, who had recently been hired, were videotaped so that they could be shown to large number of interviewers. Job performance ratings for the interviewees were collected after they had been employed at least half a year. This approach was taken because of the logistical difficulty of obtaining a sufficiently large number of interviewees who are subsequently hired and for whom job performance data is available. Moreover, there were ethical concerns about the effects that the experimental manipulations could have on applicant performance in actual interviews. However, videotaped interviews do have the advantage of allowing greater control over variables extraneous to the purpose of this study (e.g., variability among interviewees) than live interviews.

Subjects

Ninety "interviewers" (72 females and 18 males) observed and rated one of three videotaped interviews in a laboratory setting (data from two additional male subjects were not used after they indicated they had not taken the task seriously) . The interviewers were all enroled in a professional degree program in social work. All had relevant work experience and 14 were employed full-time in the field of social work at the time of the study (i.e., they were part-time students). Additionally, 25 participants had previous or current experience as interviewers.

Materials

In the present study, "predictor" data (i.e., interview ratings) are drawn from a laboratory setting whereas criterion data are drawn from a field setting. Although it would have been preferable to use a predictive validation strategy, this approach proved to be unfeasible. A very large number of applicant volunteers would have been required to ensure that at least some of them would have been hired and that, of those hired, a sufficient number would remain with the organization long enough to obtain reasonably accurate performance ratings. The administrators of the participating social services department suggested that a large number of volunteers would be very unlikely given the already high levels of stress experienced by most applicants during the interview process. Moreover, there were concerns that applicants might feel coerced into volunteering in spite of assurances that their participation in the study would have no bearing on the hiring decision.

Stimulus materials were therefore developed by videotaping employment interviews with volunteers obtained from among recently hired social welfare caseworkers employed in the social services department of a mid-sized regional government. Volunteers were recruited by promoting participation in the study as an opportunity to practice interviewing skills and receive feedback in preparation for future job opportunities (e.g., promotions). The interviews consisted of 12 behavior description (BDI) questions and an equal number of low job-relevance questions. The BDI questions were developed from a job analysis with the aid of a team of social work supervisors and administrators (e.g., "Think about the last time when you had to deal with a particularly 'difficult' individual. Describe the situation and tell me what you did."). The low job-relevance questions were those most frequently asked by interviewers in the social services department and are the kind of questions advocated in popular literature for interviewers (e.g., "What do you consider to be your strengths?", "What are your weaknesses?", "Why should we hire you?", etc.). Answers to such questions often do not provide job-relevant information.

A performance appraisal instrument, based on the same job analysis that was used to construct the BDI questions, was developed with the assistance of the social work supervisors. Job performance ratings on this instrument were obtained for each of the volunteers after they had worked at least half a year and these ratings served as the criterion data. The videotapes of one below average job performer (mean score of 3.59 on a five-point scale), one average job performer (mean score of 4.11), and one above average job performer (mean score of 4.71), relative to the volunteer group average, were selected for laboratory viewing. All three of the volunteers appearing as interviewees in the selected videotapes were female (only one male had volunteered).

Each of the selected interview videotapes was edited in order to create two separate videotapes from the original tape. One of the edited videotapes contained only questions of high job relevance (BDI) along with respective answers and the other edited videotape contained only questions of low job relevance along with respective answers. Thus, the editing process yielded six separate videotapes, each about 20 minutes in length.

Procedure

Each of the 90 subjects was randomly assigned to one of the six conditions (as described above) so that there were 15 subjects per condition. Subjects in each of the experimental conditions were asked to view one of the videotaped interviews and to rate the interviewee while assuming the role of "interviewer". As a result, five subjects in each condition saw one videotape, another five subjects saw the second videotape, and the remaining five subjects saw the third videotape. A repeated measures design in which each subject would have seen all three videotapes was not feasible because of the time constraints imposed by subjects' timetables and academic obligations. Rather than one hour, such a design would have required two to three hours of subject time. Subjects' interview ratings were correlated with supervisors' performance ratings, producing one criterion-related validity coefficient for each experimental condition.

Subjects in condition one were asked to rate answers to BDI questions as they were given using scoring guides; subjects in condition two were also asked to rate answers to BDI questions as they were given but used graphic rating scales; subjects in condition three were asked to make concise notes and rate answers to BDI questions at the conclusion of the interview using scoring guides; subjects in condition four were also asked to make concise notes and rate answers to BDI questions at the conclusion of the interview but used graphic rating scales; subjects in condition five were asked to rate answers to low job-relevance questions as they were given using graphic rating scales; and subjects in condition six were asked to make concise notes and rate answers to low job-relevance questions at the conclusion of the interview using graphic rating scales.

It should be noted that the interview scoring guides were developed prior to the videotaping of the interviews so that knowledge of the interviewees' performance in the interview would not bias the content of the scoring guide. The interviews were, in turn, videotaped prior to the collection of job performance data.

Analyses

Because each subject saw only one interview, it is not possible to assess the validity of any one subject's interview score nor can the effects of interviewer demographics variables such as amount of work experience, experience as an interviewee, experience as an interviewer and gender be partialled out. However, a MANOVA reveals that these variables were distributed relatively evenly across conditions ($\underline{F}$ (20, 336) = 1.129, $\underline{p} > .30$) and should therefore have relatively equal effects in all conditions.

The validity coefficients were obtained by correlating total interview scores (i.e., the sum of scores across individual interview questions) with total job performance ratings (i.e., the sum of ratings across individual performance items) across all 15 subjects and three videotapes in each condition. However, because predictor data is on an interval scale whereas criterion data is on an ordinal scale, Jaspen's Coefficient of Multiserial Correlation (M) was used (Freeman, 1965). M can be treated as a Pearson r (much like the Biserial Correlation Coefficient, which is a special case of M). This procedure produced one validity coefficient per condition. Total interview and performance scores were used because these scores are typically used in the final hiring or promotion decision.

Inter-rater reliability was assessed by computing a generalizability coefficient for each condition across the three interviewees based on for the total interview score given to each interviewee by the five subjects who viewed her videotaped interview (Crocker & Algina, 1986). This statistic is ideal for assessing inter-rater reliability when each candidate is rated by several raters, but each rater rates only one candidate. As with validity coefficients, only one reliability coefficient is produced per condition using this procedure.

Hypotheses one and two were tested by comparing the predicted orders of validity and reliability coefficients with the obtained orders of validity and reliability coefficients across the six conditions. Validity and reliability coefficients are expected to progressively decrease from condition one to six so that they are highest in condition one and lowest in condition six. Simple counting rules for ordered combinations (Hays, 1981, pp. 112-113) provided the probability of obtaining the predicted order. In addition, the more conventional but statistically less economical z-tests for differences between validity coefficients across the three variables hypothesized to moderate interview validity (job relevance, timing, and rating scale) were carried out after a Fisher r-to-z transformation (Hays, 1981). Differences between reliability coefficients were also tested using the UX_1 and W statistics developed by Feldt, Woodruff, and Salih (1987) and Feldt (1969), respectively, for testing differences between Cronbach's alpha coefficients (Cronbach's coefficient alpha is a special case of the generalizability coefficient). The UX_1 (approximately distributed as χ^2) is an omnibus test which can be followed with univariate W tests (approximately distributed as F) if it is significant.

Results

Table 1 presents obtained validity and reliability coefficients for each of the four experimental conditions. The obtained order of magnitude of validity coefficients across the six conditions is as predicted in hypothesis one ($p < .01$, counting rules for ordered combinations, Hays, 1981). Subjects who scored the answers to BDI questions during the interview using a scoring guide (condition one) made the best predictions. Subjects in each of the other conditions did progressively worse, from condition one to condition six, in the hypothesized order.

Insert Table 1 about here

The effect of timing was examined more conventionally by averaging z statistics (following a Fisher r-to-z transformation) across the job relevance and rating scale variables. The effect of job relevance was examined by averaging z statistics across the timing variable only for graphic rating scale conditions (2, 4, 5, and 6) because there were no scoring guides in the low job-relevance conditions. Similarly, the effect of rating scale was examined only for BDI conditions (1, 2, 3, and 4). Differences between z statistics are significant for job relevance and timing ($z = 2.03$ and $z = 1.69$, respectively) and approach significance for rating scale ($z = 1.50$). That is, validity is greater for BDI questions (mean $r = .52$) than for low-job-relevance questions (mean $r = .03$); validity is greater when answers are scored as they are given (mean $r = .62$) than when they are scored after the interview (mean $r = .34$); and validity of BDI questions tends to be greater when scoring guides are used (mean $r = .76$) than when graphic rating scales are used (mean $r = .52$). These results also lend support to hypothesis one.

Hypothesis 2 is supported as well. The generalizability coefficients are ordered in magnitude across the six conditions as predicted ($p < .01$, counting rules for ordered combinations, Hays, 1981). This order of results parallels the order obtained for validity coefficients.

As with validity, the effect of each of the three hypothesized moderating variables on inter-rater reliability was also examined by testing differences between generalizability coefficients following the omnibus test, UX_1 (5, $N = 90$) = 25.78, $p < .001$. Reliability was greater across the BDI conditions ($\hat{\rho}^2 = .44$) than across the low-job-relevance conditions ($\hat{\rho}^2 = 0.00$), W (20, 20) = 3.55, $p < .01$. Reliability was also greater when answers were scored as they were given ($\hat{\rho}^2 = .49$) than when they were scored after the interview ($\hat{\rho}^2 = .02$), W (30, 30) = 1.91, $p < .05$. Finally, reliability was greater when scoring guides were used ($\hat{\rho}^2 = .80$) than when graphic rating scales were used ($\hat{\rho}^2 = .44$), W (20, 20) = 2.85, $p < .05$. These results provide further support for hypothesis two.

Discussion

Prior to considering the results, it should be noted that the sample size in each condition is not very large. As a result the confidence interval around each of the obtained validity coefficients is fairly generous. Caution should therefore be exercised when making inferences from these coefficients. However, the purpose of this study is not to identify the precise validity of particular approaches to interviewing but, rather, to explore the relative contributions of three variables to interview validity. In other words, attention should focus on the order of magnitude of validity coefficients rather than on their absolute value.

The results of this study provide support for hypothesis one. Obtained validity coefficients were aligned in the predicted order of magnitude. These results suggest that, of the variables investigated, the job relevance of interview questions may be the most important to increasing interview validity. The timing of interview ratings and kind of rating scale used follow in decreasing order of importance.

The validity coefficient of .85 obtained in the optimal condition (one) is of particular interest. Condition one represents a highly structured interview, utilizing job-relevant questions, the answers to which are scored as they are given using a scoring guide. The obtained validity coefficient in this condition compares very favourably with the coefficients reported by Janz (1982) and Orpen (1985). In fact, the lower bound of the 95% confidence interval for the validity coefficient in condition one (.59) exceeds all but one of the coefficients reported by Janz and Orpen. The BDI used in Janz (1982) and Orpen (1985) is most similar to the interview process used in condition four of the present study (i.e., notes are taken but scoring is done after the interview using graphic rating scales). Although the validity coefficients obtained by Janz (1982) and Orpen (1985) are a little higher than that obtained in condition four (.36), they all fall within the 95% confidence interval. Differences in obtained validity coefficients may be due to the fact that Janz's and Orpen's interviewers received some training on the proper administration of the BDI whereas interviewers in the present study only received brief instructions (10 minutes). These results suggest that constructing scoring guides for the BDI and scoring answers to BDI questions as they are given may increase the validity of the BDI.

The second hypothesis in this study was that reliability coefficients would progressively decrease from condition one to six so that they would be highest in condition one and lowest in condition six. In fact, inter-rater or inter-interviewer reliability coefficients follow the same order of magnitude as validity coefficients, lending support to hypothesis two. These results were not unexpected in that reliability places an upper limit on validity (e.g., Nunnally, 1978).

As with interview validity, the job relevance of interview questions, the timing of interview ratings, and the kind of rating scale used contribute significantly, in decreasing order of importance, to the reliability of the interview. The highest reliability coefficient (.876) was also obtained in condition one where answers to job relevant questions were scored as they were given using a scoring guide. This coefficient is greater than those obtained by Janz (1982) and Orpen (1985) and reflects the very high level of inter-rater agreement achievable with the BDI if scoring guides are used and answers are scored as they are given. When such strong inter-rater agreement can be attributed to the interview instrument, board or panel interviews may provide redundant information. That is, when a given applicant receives virtually the same ratings from each interviewer on the interview board, it may only necessary to have one interviewer conduct the interview rather than an entire board. In fact, Wiesner and Cronshaw (1988) found no differences between validity coefficients for structured interviews conducted by individual interviewers and those conducted by interview boards. An employer using a structured interview could therefore offset the cost of developing the interview by using individual interviewers rather than boards to conduct interviews.

Although the results of the present investigation are promising, additional research is needed before recommendations can be made for interview practice. The highly structured interview used in the optimal experimental condition (one) should be used for selection in an applied setting in order to corroborate the validity and reliability coefficients achieved in the present study. Moreover, additional research should be conducted to investigate the information processing engaged in by the interviewer in this kind of employment interview, and how these processes differ from those occurring in the unstructured interview. Specifically, this research should focus on differences in the encoding of information and the impact of irrelevant information on the encoding process and on interview outcomes. The present study focused on the question of interview validity rather than these information processing issues.

The design used in this study has a number of advantages and disadvantages or limitations which need to be considered. Unlike the traditional validation design where a few interviewers interview a large number of interviewees, the present design employed a large number of interviewers and only three interviewees. Employment interviews were videotaped with actual, recently hired, employees who were recruited in a field setting. The resulting videotapes were shown to subjects who assumed the role of interviewer in a controlled laboratory setting. Job performance ratings, which were obtained from their supervisors, served as the criterion. This unique design was born of necessity, for situational constraints precluded the availability of a large number of interviewees. However, the design does have a number of advantages.

In the traditional validation design differences among interviewees are captured in the data collected. If the purpose of the study is to investigate how interviewers treat interview information, however, differences among interviewees contribute to error variance. In the design used in this study, variance attributable to interviewees was kept to a minimum through the use of videotapes. It was possible for many interviewers to see the same interviewee with no variation in her interview performance. This approach makes it possible, therefore, to experimentally examine the effects of different variables on interviewers' ratings, decisions, and information processing or to investigate the effects of interpersonal differences among interviewers while controlling for extraneous variables. Such control is more difficult using the traditional design.

In practical terms, it may sometimes be difficult to obtain a sufficient number of interviewees to conduct a validation study, as was the case in the present investigation. However, it may be possible recruit a large number of interviewers. In such circumstances the design used in this study provides an alternative option.

Recently hired employees, rather than job applicants, were interviewed in this study, also because of situational constraints. Although it may be argued that this design limits the generalizability of the findings somewhat, Barrett, Phillips, and Alexander (1981) found that using a concurrent rather than a predictive validation design has little practical impact on the size of the validity coefficient. Nevertheless, the same procedure could be used in the context of a predictive validation design. Interviews with volunteer job applicants could be videotaped and the videotapes for those applicants who are hired could be retained. Job performance ratings could then be collected once the applicant had been employed for a sufficient period of time.

In the present study each interviewer saw only one videotape due to time constraints. However, a repeated measures design can also be incorporated in the procedures which were used. Each interviewer could view two or more videotaped interviews and the order of presentation could be randomized. Such a design would lend itself very well to the study of contrast effects.

The methodology used in this study does have the drawback of sacrificing some of the realism of the actual interview. Because the interviewees had already been hired they might not have taken their performance in the interview seriously. It is therefore possible that their responses may have differed from what they would have been in an actual interview. However, post-interview debriefings with the interviewees suggest that this was not the case. Without exception, interviewees reported that they had tried to do as well as they could. Moreover, several of the interviewees volunteered that this interview was the most thorough and gruelling that they had ever had. Thus, based on the interviewees' reports, the fact that they were not competing for actual jobs appears to have had minimal effect on their interview performance. Moreover, post-experiment debriefings with interviewer subjects revealed than none of them suspected that the intervieews were not real applicants engaged in actual selection interviews.

However, concerns about realism for the interviewer subjects are a bit more troublesome. Even if the subjects had all been experienced interviewers (only 25 were), the artificiality of the laboratory setting may have detracted from the realism of the interview. Decisions made in a lab do not have the potential ramifications that decisions made by actual interviewers have. In order to challenge them somewhat, interviewers were told that their predictions would be compared with the actual job performance of the interviewees, should they be hired. In post-interview debriefings, all except two of the interviewers indicated that they had taken the task seriously (the data for these two were not used). In addition, the results of several studies suggest that using students as "interviewers" poses little, if any, threat to the generalizability of the results (Bernstein, Hakel, & Harlan, 1975; Dipboye, Fromkin, & Wiback, 1975; McGovern, Jones, & Morris, 1979). Bernstein et al. pointed out that students and professional interviewers do not differ in their interview ratings and decisions except that students tend to be a little more lenient in their ratings. Nevertheless, the fact that interviewer subjects were not interacting with interviewees applying for actual jobs may limit the generalizability of the study somewhat.

In conclusion, the three variables expected to contribute to the reliability and criterion-related validity of the interview (i.e., job relevance, the timing of interview ratings, and the type of rating scale used) were found to do so in the order hypothesized. In the optimal condition, the interview was constructed so that the questions were highly job-relevant and answers to interview questions were scored by the interviewer as soon as they were given using a scoring guide. Very high criterion-related validity was achieved when this carefully constructed, structured interview was used. In fact, the improved validity of the structured interview appears to be a function of its greater reliability. When the interview is structured according to the procedures described above, it can be a valid and useful selection instrument. Although there are some limitations to the generalizability of these results, they provide useful information for further research and suggestions which might be explored in practice. Moreover, the unique methodology employed in this study has some advantages which could assist future interviewing researchers in dealing with practical constraints as well as in conducting studies where they wish to isolate the effects of particular variables while retaining as much realism and workplace relevance as possible.

Further studies should be conducted in applied settings using the approach to structuring the interview employed in this study, not only to verify the validity obtained in the present study, but also to identify practical problems which may occur in field settings in the design and implementation of the structured interview. Should they be supported by further research, the results of this study suggest that interviews be constructed so that the information obtained is exclusively job related (i.e., based on a job analysis). In addition, job-performance-referenced scoring guides (cf., Latham et al., 1980) should be used to score answers as they are given by applicants. Although the development of such interview questions and scoring guides is costly, the benefits of a more effective selection interview far outweigh any expenses in most employment situations (Cascio, 1991).

References

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. Psychological Bulletin, 86, 736-765.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. Personnel Psychology, 35, 281-322.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 66, 1-6.

Bass, B. M. (1951). Situational tests:I. Individual interviews compared with leaderless group discussions. Educational and Psychological Measurement, 11, 67-75.

Bernstein, V., Hakel, M. D., & Harlan, A. (1975). The college student as interviewer: A threat to generalizability. Journal of Applied Psychology, 60, 266-268.

Bingham, W. V. D., & Moore, B. V. (1931). How to interview. New York: Harper and Brothers.

Carlson, R. E., Thayer, P. W., Mayfield, E. C., & Peterson, D. A. (1971). Improvements in the selection interview. Personnel Psychology, 50, 268-275.

Cascio, W. F. (1991). Costing Human Resources: The Financial Impact of Behavior in Organizations (3rd ed.). Boston: PWS-Kent.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. Journal of Applied Psychology, 60, 39-43.

Fear, R. A. (1958). The evaluation interview. New York: McGraw-Hill.

Fear, R. A., & Jordan, B. (1943). Employee evaluation manual for interviewers. New York: The Psychological Corporation.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363-373.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. Applied Psychological Measurement, 11, 93-103.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Freeman, L. C. (1965). Elementary Applied Statistics. New York: John Wiley & Sons, Inc.

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. Personnel Psychology, 42, 691-726.

Hays, W. H. (1981). Statistics (3rd ed.). New York: Holt, Rinehart, and Winston.

Heneman III, H. G., Schwab, D. P., Huett, D. L., & Ford, J. J. (1975). Interviewer validity as a function of interview structure, biographical data, and interviewer order. Journal of Applied Psychology, 60, 748-753.

Hovland, C. I., & Wonderlic, E. F. (1939). Prediction of industrial success from a standardized interview. Journal of Applied Psychology, 23, 537-546.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. Journal of Applied Psychology, 67, 577-580.

Janz, T., Hellervik, L., & Gilmore, D. C. (1986). Behavior description interviewing: New, accurate, cost effective. Boston, MA.: Allyn and Bacon, Inc.

Kenagy, H. G., & Yoakum, C. S. (1925). The selection and training of salesmen. New York: McGraw-Hill.

Landy, F. J. (1976). The validity of the interview in police officer selection. Journal of Applied Psychology, 61, 193-198.

Langdale, J. A., & Weitz, J. (1973). Estimating the influence of job information on interviewer agreement. Journal of Applied Psychology, 57, 23-27.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. Journal of Applied Psychology, 69, 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. Journal of Applied Psychology, 65, 422-427.

Maas, J. B. (1965). Patterned scaled expectation interview: Reliability studies on a new technique. Journal of Applied Psychology, 49, 431-433.

Mayfield, E. C. (1964). The selection interview - a re-evaluation of published research. Personnel Psychology, 17, 239-260.

Mayfield, E. C., Brown, S. H., & Hamstra, B. W. (1980). Selection interviewing in the life insurance industry: An update of research and practice. Personnel Psychology, 33, 725-739.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., Hunter, J. E., Maurer, S., & Russel, J. (1987). The validity of employment interviews: A review and meta-analysis. Unpublished manuscript.

McGovern, T. V., Jones, B. W., & Morris, S. E. (1979). Comparison of professional versus student ratings of job interviewee behavior. Journal of Counseling Psychology, 26, 176-179.

McMurry, R. N. (1947). Validating the patterned interview. Personnel, 23, 263-272.

Nunnally, J. C. (1978). Psychometric theory (2nd Edition). New York: McGraw-Hill.

Orpen, C. (1985). Patterned behavior description interviews versus unstructured interviews: A comparative validity study. Journal of Applied Psychology, 70, 774-776.

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.

Rowe, P. M. (1981). The employment interview: A valid selection procedure. Canadian Personnel and Industrial Relations Journal, 28, 37-40.

Rundquist, E. A. (1947). Development of an interview for selection purposes. In G. A. Kelly Ed., New methods in applied psychology (pp. 85-95). College Park, MD.: University of Maryland.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. Personnel Psychology, 29, 79-101.

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

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

Viteles, M. S. (1932). Industrial psychology. New York: W. W. Norton and Company.

Wagner, R. (1949). The employment interview: A critical summary.

Personnel Psychology, 2, 17-46.

Walsh, U. R. (1975). A test of the construct and predictive validity of a structured interview. Unpublished doctoral dissertation, University of Nebraska, Lincoln, NE.

Webster, E. C. (1982). The employment interview: A social judgement process. Schomberg, Ont.: S.I.P. Publications.

Wiener, Y. & Schneiderman, M. I. (1974). Use of job information as a criterion in employment decisions of interviewers. Journal of Applied Psychology, 59, 699-704.

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.

Wright, O. R. (1969). Summary of research on the selection interview since 1964. Personnel Psychology, 36, 355-370.

Yonge, K. A. (1956). The value of the interview: An orientation and a pilot study. Journal of Applied Psychology, 40, 25-31.

Table 1

Validity and Inter-rater Reliability Coefficients Across the Six Conditions

RATING SCALE	TIMING OF INTERVIEW RATINGS			
	DURING INTERVIEW		POST INTERVIEW	
	JOB RELEVANCE		JOB RELEVANCE	
	High	Low	High	Low
Scoring Guide	1 .847 (.876)	---	3 .627 (.733)	---
Graphic Rating Scale	2 .659 (.856)	5 .120 (.000)[a]	4 .358 (.030)	6 -.062 (.000)[b]

Note. Reliability coefficients are given in parentheses. For both validity and reliability coefficients, n = 15 per cell.

[a] Unadjusted generalizability coefficient = -.275

[b] Unadjusted generalizability coefficient = -.705

Faculty of Business
McMaster University

WORKING PAPERS - RECENT RELEASES

351. John P. Liefeld, Thomas E. Muller, "How Affective vs. Informative Newspaper Advertisements Bias Thoughts and Memories", August, 1990.

352. Christopher K. Bart, "Controlling New Products: Some Lessons for Success", August, 1990.

353. R.G. Cooper and R. de Brentani, "New Industrial Financial Services: What Distinguishes the Winners", September, 1990.

354. Thomas E. Muller, "Value-Based Determinants of Tourist Satisfaction Upon Visiting a Foreign City", October, 1990.

355. Thomas E. Muller, "When Americans and Canadians Visit a City: Cross-Cultural Contrasts in Sources of Tourist Satisfaction", December, 1990.

356. Robert G. Cooper, "The NewProd Model: The Industry Experience", March, 1991.

357. Min S. Basadur, "Managing the Creative Process in Organizations", April, 1991.

358. Min S. Basadur, "Impacts and Outcomes of Creativity in Organizational Settings", April, 1991.

359. Thomas E. Muller, Emmanuel J. Cheron, "Values and Ownership Patterns: Comparisons between the Canadian Provinces of Ontario and Quebec", April 1991.

360. Roy J. Adams, "Efficiency is not Enough", April, 1991.

361. Roy J. Adams, "Employment Relations in an Era of Lean Production", April, 1991.

362. John W. Medcof, John Bachynski, "Measuring Opportunities to Satisfy the Needs for Achievement, Power and Affiliation", April, 1991.

363. Robert F. Love, Paul D. Dowling, "A Nested Hull Heuristic for Symmetric Traveling Salesman Problems", June, 1991.

364. Robert F. Love, John H. Walker, and Moti L. Tiku, "Confidence Levels for $\ell_{k,p,\theta}$ Distances", June, 1991.

365. Archer, N. P., and G. O. Wesolowsky, "A Dynamic Service Quality Cost Model with Word-of-Mouth Advertising", July, 1991.

366. John W. Medcof, "Weiner, Kelley, Jones, and Davis, Peat: Integrated", August, 1991.

367. Harish C. Jain and Rick D. Hackett, "A Comparison of Employment Equity and Non-Employment Equity Organizations On Designated Group Representation And Views Towards Staffing", August, 1991.

368. J. Brimberg and R. F. Love, "Local Convergence in a Generalized Fermat-Weber Problem", August, 1991.

369. Rick D. Hackett and Peter Bycio, "The Temporal Dynamics Associated with Absenteeism as a Coping Mechanism", September, 1991.

370. Robert F. Love and John H. Walker, "A New Criterion For Evaluating the Accuracy of Distance Predicting Functions", October, 1991.

371. Joseph B. Rose, "Interest and Rights Disputes Resolution in Canada, New Zealand and Australia", October, 1991.

372. Paul A. Dion and Peter M. Banting, "Buyer Reactions to Product Stockouts in Business to Business Markets", October, 1991.

373. Isik U. Zeytinoglu, "Part-Time and Other Non-Standard Forms of Employment: Why are They Considered Appropriate for Women?", November, 1991.

374. Rick D. Hackett, Peter Bycio and Peter Hausdorf, "Further Assessments of a Three-Component Model of Organizational Commitment", January, 1992.

375. Y. Lilian Chan, Bernadette E. Lynn, "Evaluating Tangibles and Intangibles of Capital Investments", February, 1992.

376. Christopher K. Bart, "Strategic Lessons for Today's Airline CEO's", February, 1992.

377. Yufei Yuan, "A Stable Residence Exchange Problem", March, 1992.

378. Jim Gaa, "The Auditor's Role: The Philosophy and Psychology of Independence and Objectivity". April, 1992.

379. David J. Buchanan and George O. Wesolowsky, "Algorithm for Rectilinear Distance Minimax Single Facility Location in the Presence of Barriers to Travel", April, 1992.

380. Thomas E. Muller and D. Wayne Taylor, "Eco-Literacy Among Consumers: How Much Do They Know About Saving Their Planet?", April, 1992.

381. Willi H. Wiesner, "The Contributions of Job Relevance, Timing, and Rating Scale to the Validity of the Structured Employment Interview", September, 1992.