Journal of Applied Psychology
1994, Vol. 79, No. 4, 599-616

0021-9010/94/$3.00

The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis

Michael A. McDaniel, Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer

This meta-analytic review presents the findings of a project investigating the validity of the employment interview. Analyses are based on 245 coefficients derived from 86,311 individuals. Results show that interview validity depends on the content of the interview (situational, job related, or psychological), how the interview is conducted (structured vs. unstructured; board vs. individual), and the nature of the criterion (job performance, training performance, and tenure; research or administrative ratings). Situational interviews had higher validity than did job-related interviews, which, in turn, had higher validity than did psychologically based interviews. Structured interviews were found to have higher validity than unstructured interviews. Interviews showed similar validity for job performance and training performance criteria, but validity for the tenure criteria was lower.

The interview is a selection procedure designed to predict future job performance on the basis of applicants' oral responses to oral inquiries. Interviews are one of the most frequently used selection procedures, perhaps because of their intuitive appeal for hiring authorities. Ulrich and Trumbo (1965) surveyed 852 organizations and found that 99% of them used interviews as a selection tool. Managers and personnel officials, especially those who are interviewers, tend to believe that the interview is valid for predicting future job performance. In this article, we quantitatively cumulate and summarize research on the criterion-related validity of the employment interview.

Our purpose in this article is threefold. First, we summarize past narrative and quantitative reviews of criterion-related validity studies of the employment interview. Second, we report research that extends knowledge of the criterion-related validity of interviews through meta-analyses conducted on a more comprehensive database than has been available to earlier investigators. Third, we examine the criterion-related validity of different categories of interviews that vary in type and in structure.

Traditional Reviews of Interview Validity

Seven major literature reviews of interview research have been published during the past 35 years (Arvey & Campion, 1982; Harris, 1989; Mayfield, 1964; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949; O. R. Wright, 1969). Only 25 of the 106 articles on the validity or the reliability of interviews located by Wagner reported any quantitative data. Wagner's major conclusions were that (a) quantitative research on the interview is much needed; (b) the validity and reliability of the interview may be highly specific to both the situation and the interviewer; (c) the interview should be confined to evaluating factors that cannot be measured accurately by other methods; (d) the interview is most accurate when a standardized approach is used; and (e) the interviewer must be skilled in eliciting complete information from the applicant, observing significant behavior, and synthesizing developed information.

Mayfield (1964) made prescriptive statements that he considered justified by empirical findings. He concluded that typical unstructured interviews with no prior data on the interviewee are inconsistent in their coverage. He also found that interview validities are low even in studies with moderate reliabilities, although structured interviews generally show higher interrater reliabilities than do unstructured interviews. Mayfield indicated that individual interviewers, although consistent in their approach to interviewees, are inconsistent in their interpretation of data, perhaps because interviewers' attitudes bias their judgments and because there is a tendency for interviewers to make decisions early in the unstructured interview. Concerning the assessment of cognitive ability, Mayfield concluded that intelligence is the human quality that may be best estimated from an interview but that interviewer assessments offer little, if any, incremental validity over test scores alone. When the test score was known, the interview contributed nothing to the predictive validity in a multiple-assessment procedure. Mayfield also concluded that the interrater reliability of the interview was satisfactory. However, interrater reliability estimates were not based on two independent interviews but, rather, were correlations between raters in the same interview. We return to this point later.

On the other hand, Ulrich and Trumbo (1965) concluded that personal relations and motivation were two areas that contributed to most decisions made by interviewers and that these two attributes were valid predictors. They concluded that the

Michael A. McDaniel, Department of Psychology, University of Akron; Deborah L. Whetzel, American Institutes for Research, Washington, DC; Frank L. Schmidt, Department of Management and Organizations, University of Iowa; Steven D. Maurer, Department of Management, Old Dominion University.

John Hunter and James Russell made contributions to earlier versions of this article; their contributions are appreciated. We are also indebted to many authors of primary validity studies for releasing their results to us. We appreciate the substantial assistance of Allen Huffcutt in obtaining many obscurely published interview studies and for insightful critiques of drafts of the article. We also appreciate the data provided to us by Willi Wiesner and Steve Cronshaw that permitted the interstudy reliability analyses.

Correspondence concerning this article should be addressed to Michael A. McDaniel, Department of Psychology, University of Akron, Akron, Ohio 44325-4301.

use of the interview should be limited to one or two content areas as part of a sequential testing procedure.

O. R. Wright (1969) summarized the research on the selection interview since 1964 and concluded that (a) interview decisions are made on the basis of behavioral and verbal cues; (b) rapport between interviewer and interviewee is an important variable that influences the effectiveness of the interview; and (c) structured or patterned interviews are more reliable than unstructured interviews.

Schmitt (1976) reviewed the literature on interviews and found that nearly all of the recent studies on employment interviews focused on factors that influenced interview decisions rather than on the outcomes resulting from those decisions. Examples of factors include research on photographs and the effects of appearance (Carlson, 1967); contrast effects (Carlson, 1970; Hakel, Ornesorge, & Dunnette, 1970; Wexley, Yukl, Kovacs, & Sanders, 1972); situational variables, including quota position (Carlson, 1967); and race of the interviewer (Ledvinka, 1973).

Arvey and Campion (1982) reviewed the literature from 1975 to 1982 and found that there was an increase in research investigating possible bias in the interview. Attention had been focused on the interaction of group membership variables and interviewers' decision making. They also found that researchers were investigating several variables that influenced the interview, such as nonverbal behavior, interviewees' perceptions of interviewers, and interviewer training. Arvey and Campion considered most of these studies to be microanalytic in nature, studying only a narrow range of variables. They also concluded that researchers were becoming more sophisticated in their research methods and strategies but had neglected the person–perception literature, including attribution models and implicit personality theory. Finally, Arvey and Campion postulated reasons for the continued use of the interview in spite of the previous evidence of its low validity and reliability.

Most recently, Harris (1989) summarized the qualitative and quantitative reviews of interview validity and concluded that, contrary to the popular belief that interviews lack validity, recent evidence suggested that the interview had at least moderate validity. In addition, Harris addressed other issues related to the interview, such as decision making, applicant characteristics, and interviewer training.

Previous Quantitative Reviews of Interview Validity

In the past 20 years, five quantitative reviews of interview validity have been conducted (Dunnette, Arvey, & Arnold, 1971; Hunter & Hunter, 1984; Reilly & Chao, 1982; Wiesner & Cronshaw, 1988; P. M. Wright, Lichtenfels, & Pursell, 1989). The three earliest reviews may be termed *quantitative* because the mean observed validity was calculated for each set of studies. However, they were not typical validity generalization studies (Hunter & Schmidt, 1990), because estimates of the population true validities were not calculated, and variance statistics needed for validity generalization inferences were not reported. Furthermore, these studies did not examine the validity of various types of interviews separately. Hunter and Hunter's analysis included 27 coefficients analyzed separately for four criteria: supervisor ratings, promotion, training success, and tenure. Their results, shown in Table 1, indicated that the interview is a better predictor of supervisor ratings than it is of promotion, training success, and tenure; but even for supervisory ratings, the correlation was low (.14). Reilly and Chao obtained a higher validity (.19), but they had available only 12 coefficients. Dunnette et al. used 30 coefficients and obtained an average validity of .16, which is comparable to Reilly and Chao's results. Table 1 shows the number of coefficients and average validities found by each researcher, listed according to criterion.

P. M. Wright et al. (1989) conducted a meta-analysis of 13 investigations of structured interviews. This study differed from Hunter and Hunter's (1984) and Reilly and Chao's (1982) studies in that more advanced quantitative cumulation methods were used, which permitted estimation of the distribution of population validity coefficients. Initially, P. M. Wright et al. found that after correcting for criterion unreliability, the mean validity of the interview was .37. Placing the 95% credibility interval around this mean suggested that the true validity was between −.07 and .77. Because this interval included zero, they looked for moderators and outliers. After they eliminated a study reporting a negative validity coefficient (−.22; Kennedy, 1986), the mean corrected validity was .39 with a 95% credibility interval from .27 to .49. The authors noted that the interval did not include the mean validity given by Hunter and Hunter (.14; 1984) and stated that this indicated a real difference in the predictive power of the structured interview over the traditional, unstructured interview.

In the most comprehensive meta-analytic summary to date, Wiesner and Cronshaw (1988) investigated interview validity as a function of interview format (individual vs. board) and degree of structure (structured vs. unstructured). Results of this study showed that structured interviews yielded much higher mean corrected validities than did unstructured interviews (.63 vs. .20) and that structured board interviews using consensus ratings had the highest corrected validity (.64).

Need for Additional Quantitative Reviews

Wiesner and Cronshaw's (1988) meta-analysis of employment interview validities was a major integration of the literature. However, there are at least three reasons for conducting additional analyses of employment interview validity. First, the employment interview is the second most frequently used applicant-screening device (Ash, 1981). (Reviews of resumes and employment application materials are the most common; see McDaniel, Schmidt, & Hunter, 1988, for an analysis of that literature.) Given the popularity of the employment interview, more than one comprehensive review of the approach is warranted, particularly when an expanded data set such as the one used in this analysis is available. The two additional reasons for further investigation rest on a comparison between the Wiesner and Cronshaw study and the present research.

Second, the present study goes beyond Wiesner and Cronshaw's (1988) contribution to summarize validity information by criterion type. Wiesner and Cronshaw did not differentiate their criterion by job performance, training performance, and tenure. We believe that criterion-type distinctions are important because validities usually vary by criterion type, and we believe that separate analyses are warranted because of the

Table 1
Number of Validity Coefficients and Average Observed Validities Reported for the Interview by Type of Criterion

Researcher	Criterion	No. of *r*s	Mean validity
Dunnette, Arvey, & Arnold (1971)	Supervisor ratings	30	.16
Reilly & Chao (1982)	Mixture of training and performance	12	.19
Hunter & Hunter (1984)	Supervisor ratings	10	.14
	Promotion	5	.08
	Training success	9	.10
	Tenure	3	.03
Wiesner & Cronshaw (1988)	Performance (primarily supervisor ratings)	150	.26[a]
Wright, Lichtenfels, & Pursell (1989)	Supervisor ratings	13	.26[b]

[a] Disattenuated coefficient = .47. [b] Disattenuated coefficient = .39.

differences in mean reliability for the three types of criteria (Hunter & Hunter, 1984; Lilienthal & Pearlman, 1983; Pearlman, 1979; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, Pearlman, & Hirsh, 1985). We examined validity data for training performance, tenure, and job performance criteria. The validity results are reported separately, and appropriate criterion reliabilities are used.

Third, this study examines the validity of three different types of interview content: situational, job related, and psychological. Wiesner and Cronshaw (1988) did not differentiate among categories of interview content. We believe that it is likely that validities vary as a function of the content of the information collected in the interview.

In summary, although Wiesner and Cronshaw's (1988) meta-analysis was a major contribution, the present study provides a significant incremental contribution to knowledge of the interview in several areas. First, an extensive literature search was conducted that resulted in a larger number of validity coefficients than had been examined by previous researchers. Second, the validity of the interview was quantitatively assessed as a function of both interview content and structure. Third, the validity of the interview was summarized for three categories of criteria: job performance, training performance, and tenure.

We hypothesized that the validity of the employment interview depends on three factors: (a) the content of information collected during the interview, (b) how interview information is collected, and (c) criteria used to validate the interview. Each of these factors is described below.

Content of Information Collected

The first factor likely to affect interview validity is the content of the interview. In the present research, a distinction was drawn between situational, job-related (but not primarily situational), and psychological interviews. Questions in situational interviews (Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980) focus on the individual's ability to project what his or her behavior would be in a given situation. For example, an interviewee may be evaluated on the choice between winning a coveted award for lowering costs and helping a coworker who will make a significant profit for the company (Latham, 1989). Job-related interviews are those in which the interviewer is a personnel officer or hiring authority and the questions attempt to assess past behaviors and job-related information, but most questions are not considered situational. Psychological interviews are conducted by a psychologist, and the questions are intended to assess personal traits, such as dependability. Psychological interviews are typically included in individual assessments (Ryan & Sackett, 1989) of personnel that are conducted by consulting psychologists or management consultants.

Recently, some researchers have asserted the benefits of behavioral interviews (Janz, 1982, 1989). There were too few validity coefficients for these studies to be analyzed separately. Behavioral interviews describe a situation and ask respondents how they have behaved in the past in such a situation. One might argue that because the applicant is responding to a question about behavior in a specific situation, the behavioral interviews should be classified as *situational.* Others might argue that because the interviews solicit information about past behavior, the interviews should be classified as *job related.* In this study, we assigned the studies to the job-related category. As more validity data are accumulated, we recommend that the validity of the behavioral interview be evaluated as a distinct category of interview content.

These three categories were not completely disjointed. Situational interviews are designed to be job related, but they represent a distinct subcategory worthy of separate analysis. Nonpsychologist interviewers are interested in psychological traits such as dependability, and psychologists are interested in an applicant's past work experience. Also, job-related interviews, which are not primarily situational interviews, often will include some situational questions. Still, some interviews are primarily situational, others are primarily nonsituational and job related in content, and others are primarily psychological. We hypothesized that situational interviews would prove to be more valid than job-related interviews, which, in turn, we expected to be more valid than psychological interviews.

How Information Is Collected

Interviews can be differentiated by the extent of their standardization. For example, most oral board examinations con-

ducted by civil service agencies in the public sector were considered to be structured because the questions and acceptable responses were specified in advance and the responses were rated for appropriateness of content. McMurry's (1947) "patterned interview" is an early example of a structured interview. It is described as a printed form containing specific items to be covered and providing a uniform method of recording information and rating the interviewees' qualifications.

Unstructured interviews gather applicant information in a less systematic manner than do structured interviews. Although the questions may be specified in advance, they usually are not, and there is seldom a formalized scoring guide. Also, all persons being interviewed are not typically asked the same questions. This difference in interview structure may affect the interview evaluations in two ways. First, the lack of standardization may cause the unstructured interviews to be less reliable than the structured interviews. Second, structured interviews may be better than unstructured interviews for obtaining a wide range of applicant information. On the basis of these differences, we hypothesized higher validities for structured interviews.

The distinction between structured and unstructured interviews is not independent from interview content. Although some job-related interviews were structured and others unstructured, most of the psychological interviews were categorized as *unstructured,* and all of the situational interviews were classified as *structured.*

Another consideration for users of the employment interview is the use of board interviews (interviews with more than one rater in the same interview) versus individual interviews (one rater per interview). Because board interviews have higher administrative costs than individual interviews, they must be more valid if they are to be cost-effective. Several researchers (Arvey & Campion, 1982; Pursell, Campion, & Gaylord, 1980) have suggested that board interviews may be more valid than individual interviews. Wiesner and Cronshaw (1988) have provided the most extensive analysis of the validity of these two types of interviews. For unstructured interviews, board interviews were more valid than individual interviews (mean observed coefficients = .21 vs. .11, respectively). However, for structured interviews, board interviews and individual interviews were similar in validity (mean observed coefficients = .33 vs. .35, respectively). We made no hypotheses concerning differences in validity between board and individual interviews.

Another issue to consider is the interviewers' access to cognitive ability test scores. If interviewers have access to test scores, the final interview evaluations should have higher validities than if the interviewer did not have this information. This would be expected because the interviewer would have access to more criterion-relevant information.

Criteria Used to Validate the Interview

In past validity generalization studies (Hirsh, Northrop, & Schmidt, 1986; Hunter & Hunter, 1984; Lilienthal & Pearlman, 1983; Pearlman, 1979; Pearlman et al., 1980), validity has been found to vary depending on whether job performance measures or training performance measures served as the criteria. Therefore, we examined the validity of the interview for both of these criteria as well as for tenure. No specific hypotheses were made concerning the differences in validities between the criteria.

When possible, we also drew distinctions between job performance criteria collected for research purposes and for administrative purposes (i.e., performance ratings collected as part of an organizationwide performance appraisal process). This distinction was based on the premise that non-performance-related variables, such as company policies or economic factors, are more likely to influence criteria collected for administrative purposes than are those collected for research purposes. Wherry and Bartlett (1982) hypothesized that ratings collected solely for research purposes would be more accurate than ratings collected for administrative purposes. Several studies have demonstrated that ratings collected for administrative purposes are significantly more lenient and exhibit more halo than do ratings collected for research purposes (Sharon & Bartlett, 1969; Taylor & Wherry, 1951; Veres, Field, & Boyles, 1983; Warmke & Billings, 1979). Zedeck and Cascio (1982) also found that the purpose of the rating was an important correlate of ratings of performance. In the present study, all validity coefficients for training and tenure criteria were judged to be administrative criteria. We hypothesized that validities would be higher for research criteria than for administrative criteria.

Method

Meta-Analysis as a Method of Determining Validity

We used Hunter and Schmidt's (1990, p. 185) psychometric meta-analytic procedure to test the proposed hypotheses. This statistical technique estimates how much of the observed variation in results across studies is due to statistical and methodological artifacts rather than to substantive differences in underlying population relationships. Some of these artifacts also reduce the correlations below their true (e.g., population) values. Through this method, the variance attributable to sampling error and to differences between studies in reliability and range restriction is determined, and that amount is subtracted from the total amount of variation, yielding estimates of the true variation across studies and of the true average correlation. We used artifact distribution meta-analysis with the interactive method (Hunter & Schmidt, 1990, chap. 4). The mean observed correlation was used in the sampling error variance formula (Hunter & Schmidt, 1990, pp. 208–210; Law, Schmidt, & Hunter, 1994; Schmidt et al., 1993). The computer program that we used is described in McDaniel (1986). Additional detail on the program is presented in Appendix B of McDaniel et al. (1988).

In our analyses, we corrected the mean observed validity for mean attenuation due to range restriction and criterion unreliability (Hunter & Schmidt, 1990, p. 165). Although employment interviews used in selection have less than perfect reliability, the mean validity was not corrected for predictor unreliability because the goal was to estimate the operational validity of interviews for selection purposes. However, the observed variance of validities was corrected for variation across studies in interview unreliabilities. In addition, the variance of the observed validities was corrected for variation across studies in criterion unreliabilities and range-restriction values. For comparison purposes, we also reported analyses in which no range-restriction corrections were made in either the mean or variance of the interview distributions. The rationale for conducting the analyses with and without range-restriction corrections is detailed later under *Artifact Information.*

Literature Review

We conducted a thorough search for validity studies, starting with the database of validity coefficients collected by the U.S. Office of Personnel

Management (Dye, 1988). In addition, we examined references from the five literature reviews cited above to determine if the articles contained validity coefficients. Although this literature search yielded more validity coefficients than had been assembled by other reviewers (Dunnette et al., 1971; Hunter & Hunter, 1984; Reilly & Chao, 1982; Wiesner & Cronshaw, 1988; P. M. Wright et al., 1989), it is likely that we did not obtain all existing validity coefficients. However, this search, extending over a period of 8 years, is perhaps the most complete search ever conducted for interview validities. The references for the studies are listed in the Appendix.

Decision Rules

We used certain decision rules for selecting relevant studies for this meta-analysis. First, only studies using measures of overall job performance, training performance, or tenure as criteria were included. For example, Rimland's (1958) study was not included in this analysis because an attitude scale, the Career Intention Questionnaire, was the only criterion. Another study excluded was Tupes's (1950), in which the criterion was the combined judgment of three clinicians who studied each subject for 7 days and made final ratings of psychological material gathered during the assessment. Also excluded were studies in which the interview attempted to predict intelligence test scores and not job performance, training performance, or tenure. Data from Putney (1947) were excluded because the criterion contrasted the training performance of those selected by an interview with those who were not screened, and we did not consider this an interview validity coefficient.

When the criterion was on-the-job training, the distinction between job performance and training performance was not clear. For purposes of this meta-analysis, on-the-job training—when the employee was performing the job but was considered to be in training—was coded as job performance. On the other hand, results of classroom or other formal instruction were uniformly coded as training performance.

We also omitted studies in which there was no actual interview, as the interview is traditionally understood. For example, the findings of Campbell, Otis, Liske, and Prien (1962) were excluded because they reported the validity of a summary report made by a person who had access to interview scores assigned by someone else. Similarly excluded were the studies by Grant and Bray (1969) and by Hilton, Bolin, Parker, Taylor, and Walker (1955), in which the raters did not conduct the actual interviews but derived scores by reviewing narrative summaries of interview reports. We also excluded the findings of Trankell (1959), who used standardized tests to measure such traits as panic resistance and sensitivity; those of Dicken and Black (1965), who reported the validity of clinical interpretations of an objective test battery; and those of Denton (1964), who studied "interview-type data" that consisted of written responses to questions. Data from Miles, Wilkins, Lester, and Hutchens (1946) were excluded because three individuals were screened per minute. Data based on enlisted military personnel presented in Bartlett (1950) were not included because most of the interviews lasted for only a few minutes.

For our purposes, the individuals being interviewed were required to be employees or applicants for a job. We excluded data from Anderson (1954) because those interviewed were doctoral candidates. Data from Dann and Abrahams (1970), Newmann and Abrahams (1979), Walsh (1975), and Zaccaria et al. (1956) were excluded because the interviewees were college students. Psychiatric interviews were also omitted from this analysis when the interviewee was not a job applicant or employee. Examples of excluded studies include Matarazzo, Matarazzo, Saslow, and Phillips (1958) and Hunt, Herrmann, and Noble (1957), which involved psychiatric interviews conducted at mental hospitals.

The same data were often reported in more than one study. For example, we did not include data from Felix, Cameron, Bobbitt, and Newman (1945) because these data were duplicated in Bobbitt and Newman (1944). We did not code data from Johnson (1979) because an expanded data set was available in Johnson and McDaniel (1981). We did not include data from Drucker (1957) because the same data appeared to be reported in Parrish, Klieger, and Drucker (1955) and in Rundquist (1947). We did not include data from Reeb (1968) because it appeared that the same data were reported in Reeb (1969). Finally, we did not include data from Glaser, Schwartz, and Flanagan (1956) because these data were duplicated in Glaser, Schwartz, and Flanagan (1958).

Consistent with the decision rules of past validity generalization studies (Rothstein & McDaniel, 1989), we did not include incomplete data from sparse matrices. For example, studies reporting only significant correlations were omitted because capitalization on chance would cause these data to bias the mean validities upward.

If the interview was conducted as part of an assessment center or a similar multiple assessment procedure (so that the ratings could have been affected by performance in other exercises), the data were excluded from our analyses. This eliminated data from several studies (e.g., Dunnette, 1970), including a set of studies based on data obtained from the British civil service system (Anstey, 1966, 1977; Castle & Garforth, 1951; Gardner & Williams, 1973; Handyside & Duncan, 1954; Vernon, 1950; Wilson, 1948). These civil service studies incorporated the interview into assessment-center-like screening that sometimes lasted up to a week (Gardner & Williams, 1973). Note that the data from Handyside and Duncan were excluded because only corrected coefficients were reported and insufficient data were described to permit estimation of the uncorrected coefficients.

Kennedy (1986) reported several studies that contained two-part interviews. One of the parts was classified as structured by our decision rules and the other as situational. In two of the studies, validities were reported separately for each part. In these cases, we assigned one coefficient to the structured category and the other to the situational category.

We coded the retained studies by using guidelines established for a larger validity generalization project (McDaniel & Reck, 1985). Content of the interview was categorized into three groups: situational, job related, and psychological. Information on how the interview was conducted was summarized in three sets of categories: (a) structured and unstructured, (b) board and individual, and (c) test information available to the interviewer and no test information available to the interviewer. Information on the nature of the criterion was summarized into two sets of categories: (a) criterion content (job performance, training performance, or tenure) and (b) purpose for which the criterion was collected (administrative or research). Because these categories could be correlated, we used a hierarchical moderator analysis to assess the potential confounding of analyses due to correlation and interaction. In a fully hierarchical moderator analysis, the data set of correlations is broken down by one key moderator variable first, and then, within each subgroup, subsequent moderator analyses are undertaken one by one in a hierarchical manner (Hunter & Schmidt, 1990, p. 527). Although the coding scheme permitted a variety of detailed analyses, there were far too few interview validity coefficients to conduct a full hierarchical analysis. Partial hierarchical analyses were conducted when the number of available coefficients permitted a meaningful analysis. For example, hierarchical analyses were most fully implemented in studies using job performance criteria, much less so in studies using training performance criteria, and not at all with validities using tenure criteria.

When more than one coefficient from a sample was available, we invoked decision rules to include only one of the coefficients. A common reason for multiple coefficients from a sample was that more than one job performance criterion, each using a different measurement method (e.g., supervisory ratings or production data), was available. In this circumstance, the coefficient using a supervisory rating was retained. In some studies, coefficients using more than one training performance criterion, each using a different measurement method (e.g., instructor ratings or job knowledge test), were available. In this case, we retained

the coefficient based on the job knowledge test criterion. When the choice was between validities based on a single criterion measurement or on a composite measurement (e.g., a composite criterion based on written test scores and an instructor rating), we retained the coefficient based on the composite criterion. When multiple coefficients were due solely to differences in the amount of time between the collection of the predictor and criterion data, coefficients based on longer time periods were chosen over those based on shorter or unknown time periods. Interviews in which the interviewer had access to test scores were preferred. Interviews conducted by one interviewer were chosen over those conducted by multiple interviewers. The above decision rules, involving choices among criterion measurement methods, enhance the generalization of the results to the most frequently occurring criterion measurement methods and are consistent with the distributions of criterion reliabilities used in this analysis.

Reliability of Coding

When coding studies for validity generalization analyses, the reliability of interest is that between coders of studies used in the meta-analysis. Previous research (Whetzel & McDaniel, 1988) has indicated that such coding was very reliable and that independent investigators, coding the same data, record the same values and obtain the same results. In the present study, we coded data from 59 studies in common with Wiesner and Cronshaw's (1988) study. The correlation between these two data sets was .89 for the correlation coefficients and .97 for the sample sizes. A more compelling comparison of the coding reliabilities examines the meta-analytic results conducted separately on the coefficients coded from these two sets of data. When we corrected only for sampling error, the results indicated that the two sets yielded very similar mean observed coefficients (.29 vs. .26) and standard deviations (.142 vs. .164). The lack of perfect agreement in the coding of correlation coefficients can be attributed to differences in coding decision rules. For example, when separate validity coefficients were available for each dimension of an interview (e.g., interpersonal skills or job knowledge), Wiesner and Cronshaw coded the coefficient for the overall or summary dimension. We coded the coefficient between a composite of the individual dimensions and the criterion. These results support the assertion that the data in the present study were accurately coded.

Artifact Information

The studies contained little information on the reliability of the job performance and job training criteria. Therefore, the criterion reliability distributions used by Pearlman (1979) were used in this study (average criterion reliabilities of .60 and .80 for job performance and training criteria, respectively). A reviewer of this article asserted that our use of a mean criterion reliability of .60 overcorrected the observed coefficients because for some studies the criterion consisted of a composite of ratings supplied by more than one supervisor. Whereas we suspected that the concern of the reviewer may be shared by others, we sought to fully address this issue.

We cannot agree that the mean reliability of .60 for job performance ratings is an underestimate. Rothstein (1990) found that across 9,975 employees and across all time periods of supervisory exposure to employees, the mean interrater agreement (reliability for one rater) was .48. The mean of .60 applies to situations in which supervisors had about 20 years to observe employee performance. Even if we round the .48 figure up to .50, we find that by using a mean reliability of .60, we have allowed for the use of two raters in 62.5% of the studies. This is undoubtedly a much higher percentage of the studies than, in fact, had two raters. That is, the .60 figure is indeed almost certainly an overestimate, and therefore the reliability corrections were undercorrections. Below we present our analysis supporting the figure of 62.5%. Note the following: $.50x + .66(1 - x) = .60$. The .60 is the mean reliability figure we used; x is the proportion of studies in which we have (in effect) assumed there is only 1 rater (reliability = .50); and $1 - x$ is the proportion of studies in which we have (by assuming that mean reliability = .60) assumed there were two raters. The .66 is the reliability of ratings based on two raters, as determined by the Spearman–Brown formula. Solving for x finds $x = .375$, and $1 - x = .625$. Thus, on the basis of Rothstein's findings, our mean of .60 assumes that one rater was used in 37.5% of the studies and two raters were used in 62.5% of the studies. On the basis of our experience and our reading of the literature, we believe only a minority of studies use ratings by two raters. Thus, we have actually overestimated the reliability of job performance ratings.

No estimates of the reliability of tenure were available, leading, by default, to operational estimates of 1.00; this assumption of perfect reliability leads to conservative (lower bound) validity estimates for the criterion of tenure.

We reviewed the literature for studies containing information on the reliability of interviews. Some reliability coefficients were obtained from studies that reported validity coefficients. Most of the reliability coefficients, however, came from studies that did not report validity data. The desired interview reliability coefficient is obtained when applicants are interviewed twice, each time by a different interviewer, and the two evaluations are correlated. Such a reliability coefficient captures two types of measurement error: error due to the applicant giving different answers in the two interviews (temporal instability) and error due to the interviewers not agreeing in their ratings of the applicant (scorer or conspect unreliability). However, the literature (remarkably) contained only one coefficient of this type: a value of .64 found by Motowidlo et al. (1992), based on 37 interviewees. All other reliability estimates found represented interviewer reliability: The applicants participated in one interview with two or more interview raters present, and these evaluations were correlated. Thus, these reliability coefficients do not reflect the measurement error represented by applicant changes in behavior or answers in different interviews (temporal instability). Therefore, these interviewer reliability coefficients overestimate the reliability of the interview.

This is not a problem in the present analysis because the corrections to the validity distributions depend on the variance of the predictor reliability distribution, not on its mean. That is, validities were not corrected for attenuation due to interview unreliability; the only correction was for variance in validities across studies due to variability in interview reliabilities, in accordance with standard practice in validity generalization studies. Enough data were available to assemble reliability distributions for psychological, structured job-related, and unstructured job-related interviews. Descriptive statistics for these distributions are shown in Table 2. The mean reliability of the unstructured

Table 2
Descriptive Statistics for the Interviewer Reliability Distributions

Reliability distribution	No. of coefficients	Reliability		
		M	*Mdn*	*SD*
Distributions from the literature				
Psychological	25	.73	.74	.11
Job related unstructured	20	.68	.78	.25
Job related structured	167	.84	.89	.15
Pooled distributions				
All job related	187	.82	.87	.17
All reliabilities	212	.81	.87	.17

distribution was .68, whereas the mean reliability of structured interviews was .84. The pooled distribution labeled *all job related* is the combination of unstructured and structured job-related distributions. The distribution labeled *all reliabilities* is the combination of the psychological and the two job-related distributions.[1]

Most observed validity coefficients are attenuated because of range restriction in the predictor. This restriction results from employees being selected on the basis of the predictor or on the basis of another selection method that is correlated with the predictor. Range restriction is indexed as $u = s/S$, where s is the restricted standard deviation and S is the unrestricted standard deviation. Although information on the level of range restriction in employment interviews was not provided in most studies, we obtained 15 estimates of range-restriction information. At the suggestion of a reviewer, we dropped an outlier in our range-restriction distribution, reducing our distribution of range-restriction statistics to 14 observations. These data are presented in Table 3. The mean value of .68 and the standard deviation of .16 for these u values are very similar to those from validity studies of aptitude and ability tests (Alexander, Carson, Alliger, & Cronshaw, 1989; Schmidt et al., 1985). They are also very similar to the interview range-restriction data reported by Huffcutt (1992). Although the range-restriction data appear reasonable, the distribution is based on only 14 observations and, thus, might not be representative of the entire literature. Therefore, we conducted all analyses twice, once with range-restriction corrections and once without. The validity values obtained without correction for range restriction, however, must be regarded as lower bound values (i.e., downwardly biased).

Results

The results are organized to address three categories of characteristics that might covary with interview validity. Each category contains one or more subcategories of characteristics: (a) content of the interview—situational versus job-related (nonsituational) versus psychological; (b) how the interview is conducted—structured versus unstructured, board versus individual, test information available to the interviewer versus no test information available; (c) the nature of the criterion—job performance versus training performance versus tenure and administrative versus research purpose for collection of the criterion.

Table 3
Range-Restriction Information for the Interview

s/S	Frequency
.452	1
.491	1
.494	1
.505	1
.616	1
.651	1
.686	1
.689	1
.696	1
.828	1
.829	1
.830	1
.833	1
.962	1

Note. Mean s/S = .68. s = restricted standard deviation; S = unrestricted standard deviation.

The administrative versus research purpose analyses were conducted solely for the job performance criteria because all training and tenure criteria were classified as administrative criteria.

The first column of data in each table identifies the distribution of validities analyzed. The next four columns present the total sample size, the number of validity coefficients on which each distribution was based, and the uncorrected mean and standard deviation of each distribution. The next three columns present the estimated population mean (ρ), the estimated population standard deviation (σ_ρ), and the 90% credibility value for the distribution of true validities. The population distribution estimates are for distributions in which the mean true validities are corrected only for unreliability in the criterion, not predictor unreliability. We corrected the variances of the true validity distributions for sampling error and for differences among the studies in predictor and criterion reliability. Data in the last three columns of Tables 4-8 are the results of range-restriction corrections being added to the corrections discussed above. Thus, in these last three columns, the mean true validity is corrected for unreliability in the criterion and range restriction, and the variance is corrected for sampling error and for differences among studies in predictor reliability, criterion reliability, and range restriction.

The results from analyses that include range-restriction corrections were the most accurate estimates of the population validity distributions. The results based on analyses not including range-restriction corrections yielded mean validity estimates that were lower bound (downwardly biased) estimates of validity. Therefore, in the following text, we focus our discussion on the results that included range-restriction corrections.

Table 4 shows results detailing how interview validity for job performance criteria covaries with interview content, interview structure, and purpose for which the criteria are collected. Table 5 shows the same analyses for training performance criteria. Table 6 shows individual and board interviews for job performance criteria. Table 7 allows a comparison of the validity of interviews for job performance criteria where the interviewer

[1] After this article was accepted, J. Conway (personal communication, March 15, 1994) alerted us to a group of reliability coefficients obtained under conditions such that each applicant was interviewed first by one interviewer and then, on another occasion, was interviewed by a second interviewer. The mean of the 41 reliability coefficients was .52, a much lower value than the .81 average in Table 2. This mean difference has no implications for the results of this meta-analysis because, as noted before, we do not correct for mean predictor unreliability. However, it is of interest from a more general viewpoint. This finding means that, on average, 29% of the variance in interview scores—(.81 − .52) × 100—is due to transient error, that is, to occasion-to-occasion instability in applicant responses. Only 19% of the variance—(1.00 − .81) × 100—is due to disagreement between interviewers observing the same interview at a single time. That is, only 19% is due to conspect unreliability. The standard deviation of Conway's reliability coefficients (.23) was larger than that reported in our Table 2 (.17). This means that our analysis can be expected to somewhat underestimate the variance in validities that is due to variability in predictor reliabilities, resulting in a slight overestimate of σ_ρ values and a corresponding underestimation of the generalizability of interview validities. Thus, the effect is to make our results slightly conservative.

Table 4
Analysis Results for Employment Interviews in Which the Criterion Was Job Performance: Validities of Interviews by Interview Content, Structure, and Criterion Purpose

Interview distribution	N	No. rs	Mean r	Obs. σ	Without range-restriction corrections ρ	σ_ρ	90% CV	With range-restriction corrections ρ	σ_ρ	90% CV
All interviews	25,244	160	.20	.15	.26	.17	.04	.37	.23	.08
Interview content										
Situational	946	16	.27	.14	.35	.08	.26	.50	.05	.43
Job related	20,957	127	.21	.16	.28	.18	.05	.39	.24	.08
Psychological	1,381	14	.15	.10	.20	.03	.16	.29	.00	.29
Interview structure										
Structured	12,847	106	.24	.18	.31	.20	.05	.44	.27	.09
Unstructured	9,330	39	.18	.11	.23	.10	.10	.33	.12	.17
Criterion purpose										
Administrative	15,376	90	.19	.15	.25	.16	.04	.36	.22	.08
Research	5,047	65	.25	.19	.33	.20	.08	.47	.26	.13
Job-Related Interviews × Structure										
Structured	11,801	89	.24	.18	.31	.21	.04	.44	.28	.07
Unstructured	8,985	34	.18	.11	.23	.10	.10	.33	.13	.17
Job-Related Interviews × Criterion Purpose										
Administrative	12,414	74	.21	.15	.27	.17	.06	.39	.23	.09
Research	3,771	49	.27	.20	.36	.21	.08	.50	.28	.14
Job-Related Structured Interviews × Criterion Purpose										
Administrative	8,155	50	.20	.16	.26	.18	.04	.37	.24	.06
Research	3,069	36	.28	.21	.37	.23	.07	.51	.31	.11
Job-Related Unstructured Interviews × Criterion Purpose										
Administrative	4,259	24	.22	.14	.29	.13	.12	.41	.17	.19
Research	531	9	.21	.13	.27	.00	.27	.38	.00	.38

Note. Obs. = observed; ρ = estimated population mean; σ_ρ = estimated standard deviation; 90% CV = 90% credibility value for the distribution of true validities.

had access to cognitive test scores with those interviews where interviewers did not have access to cognitive test information. Table 8 shows the validity of the interview for three criteria: job performance, training performance, and tenure.

Discussion

This discussion is organized to address three categories of analyses examining characteristics that are likely to covary with

Table 5
Analysis Results for Employment Interviews in Which the Criterion Was Training Performance: Validities of Interviews by Interview Content and Structure

Interview distribution	N	No. rs	Mean r	Obs. σ	Without range-restriction corrections ρ	σ_ρ	90% CV	With range-restriction corrections ρ	σ_ρ	90% CV
All interviews	59,844	75	.23	.09	.26	.09	.14	.36	.09	.24
Interview content										
Job related[a]	51,152	56	.22	.08	.25	.08	.14	.36	.09	.24
Psychological	8,376	15	.25	.11	.28	.11	.14	.40	.14	.22
Job-Related Interviews × Structure										
Structured	3,576	26	.21	.12	.24	.09	.12	.34	.11	.20
Unstructured[a]	47,576	30	.23	.08	.25	.06	.17	.36	.05	.29

Note. Obs. = observed; ρ = estimated population mean; σ_ρ = estimated standard deviation; 90% CV = 90% credibility value for the distribution of true validities.
[a] One coefficient in this distribution, obtained from Bloom & Brundage (1947), had a sample size of 37,862 and an observed validity coefficient of .22.

Table 6
Analysis Results for Board and Individual Employment Interviews in Which the Criterion Was Job Performance

Interview distribution	N	No. rs	Mean r	Obs. σ	Without range-restriction corrections ρ	σ_ρ	90% CV	With range-restriction corrections ρ	σ_ρ	90% CV
All interviews										
Individual interviewer	11,393	90	.24	.18	.31	.20	.05	.43	.26	.09
Board interview	11,915	54	.17	.12	.22	.13	.06	.32	.17	.11
Individual Interviews × Structure										
Structured	8,944	61	.25	.19	.33	.22	.05	.46	.29	.08
Unstructured	1,667	19	.18	.11	.24	.00	.24	.34	.00	.34
Board Interviews × Structure										
Structured	2,785	35	.20	.02	.26	.13	.09	.38	.17	.15
Unstructured	6,961	15	.18	.11	.23	.12	.08	.33	.16	.13

Note. Obs. = observed; ρ = estimated population mean; σ_ρ = estimated standard deviation; 90% CV = 90% credibility value for the distribution of true validities.

interview validity: content of the interview, how the interview is conducted, and the nature of the criterion.

Content of the Interview

The analyses focusing on interview content examined interview validity across three types of interview content: situational, job related, and psychological. For job performance criteria (Table 4), situational interviews yield a higher mean validity (.50) than do job-related interviews (.39), which yield a higher mean validity than do psychological interviews (.29). For training data (Table 5), the mean validity of job-related interviews (.36) is somewhat lower than the mean validity of psychological interviews (.40).

How the Interview Is Conducted

The analyses of whether validity varies with how the interview is conducted addressed three questions: Are structured interviews more valid than unstructured interviews? Are board interviews more valid than individual interviews? and Are interviews more valid when the interviewer has access to cognitive test scores? Each of these questions is addressed in turn.

Are structured interviews more valid than unstructured in-

Table 7
Analysis Results for Employment Interviews in Which the Criterion Was Job Performance: Effects of Test Information on Validities

Test available?	N	No. rs	Mean r	Obs. σ	Without range-restriction corrections ρ	σ_ρ	90% CV	With range-restriction corrections ρ	σ_ρ	90% CV
Yes	2,196	19	.14	.11	.18	.06	.10	.26	.07	.17
No	6,843	47	.25	.19	.32	.21	.04	.45	.29	.08
Unknown	16,205	94	.19	.14	.25	.15	.05	.35	.20	.09
Structured interviews										
Yes	1,031	9	.09	.11	.11	.08	.01	.16	.11	.02
No	4,865	36	.22	.20	.29	.23	−.01	.40	.32	−.01
Unknown	6,951	61	.27	.16	.35	.17	.14	.50	.22	.22
Unstructured interviews										
Yes	433	5	.18	.06	.24	.00	.24	.34	.00	.34
No	1,854	9	.32	.12	.41	.07	.31	.57	.00	.57
Unknown	7,043	25	.14	.07	.18	.02	.16	.26	.08	.26

Note. Obs. = observed; ρ = estimated population mean; σ_ρ = estimated standard deviation; 90% CV = 90% credibility value for the distribution of true validities.

Table 8
Analysis Results for Employment Interviews: Comparison of Criterion Categories

Criteria distribution	N	No. rs	Mean r	Obs. σ	Without range-restriction corrections ρ	σ_ρ	90% CV	With range-restriction corrections ρ	σ_ρ	90% CV
Job performance	25,244	160	.20	.15	.26	.17	.05	.37	.23	.08
Training performance	59,844	75	.23	.09	.26	.09	.14	.36	.09	.24
Tenure	1,223	10	.12	.17	.13	.17	−.08	.20	.24	−.11

Note. Obs. = observed; ρ = estimated population mean; σ_ρ = estimated standard deviation; 90% CV = 90% credibility value for the distribution of the true validities.

terviews? Structured interviews, regardless of content, are more valid (.44) than unstructured interviews (.33) for predicting job performance criteria (Table 4). When the content of the interview is job related, structured interviews are still more valid (.44) than unstructured interviews (.33). However, when the criterion is training performance, the validity of unstructured and structured interviews is similar (.34 and .36, respectively), as shown in Table 5.

It should be emphasized that to obtain a correlation between interviews and a criterion, interviewers must use a rating instrument, even for unstructured interviews. Therefore, it is likely that the unstructured interviews included in this review were more structured than those typically conducted in applied settings. This suggests that the validity of most unstructured interviews used in practice may be lower than the validity found in this study. However, to the extent that unstructured interviews resemble those studied here, their validity should approximate the present results.

Are board interviews more valid than individual interviews? In board interviews, multiple interviewers provide ratings in one setting (Warmke & Weston, 1992; Weston & Warmke, 1988). Although board interviews typically are more costly than individual interviews (because they are more labor-intensive), board reviews are likely to be more reliable because more than one person provides ratings. Table 6 shows results of meta-analyses comparing the validity of individual and board interviews for job performance criteria. When all interviews are considered together, individual interviews appear more valid than board interviews (.43 vs. .32). When interviews are differentiated by structure, the results are similar. Individual interviews are more valid both when they are structured (.46 vs. .38) and when they are unstructured (.34 vs. .33).

Are interviews more valid when the interviewer has access to cognitive test scores? Table 7 shows results of analyses comparing the validity of interviews in which interviewers had access to cognitive ability test scores with the validity of interviews where no test data were available. Results showed that an interviewer's access to test scores appears to decrease validity for predicting job performance. Additional analyses showed that this is true for both structured and unstructured interviews. However, for most coefficients, it was unknown whether the interviewer had access to test scores. We suggest that conclusions are tentative and await future research.

The Nature of the Criterion

Analyses concerning the influence of the nature of the criterion on the validity of the interview focused on two questions: Does validity vary between job performance, training performance, and tenure criteria? and Does validity vary as a function of whether the criteria are collected for administrative or research purposes? Each of these questions is addressed in turn.

Does validity vary between job performance, training performance, and tenure criteria? A comparison of Tables 4 and 5 indicates that interviews are similar in predictive accuracy for job performance (.37) and training performance (.36). This pattern of validities is in contrast with cognitive ability research showing that training performance is more highly predicted than are supervisory ratings of job performance (e.g., Lilienthal & Pearlman, 1983; Pearlman, 1979). However, when job performance is measured with content-valid job sample tests, cognitive aptitude and ability tests reveal a pattern similar to that noted here for interviews, in that they have been found equally valid for job performance and training performance criteria (Schmidt et al., 1985; Schmidt, Ones, & Hunter, 1992). Table 8 shows the validity of the interview for job performance, training performance, and tenure. This table indicates that tenure is less predictable by the interview (.20) than are job and training performance. In the absence of a feasible method for assessing the reliability of tenure, it was assumed by default to be perfect, leading to a lower bound validity estimate. As in other analyses involving few coefficients, the tenure results are best viewed as tentative pending further study.

Does validity vary as a function of whether the criteria were collected for administrative or research purposes? Job performance criteria may be collected for administrative purposes or for research purposes only. As shown in Table 4, the mean validity with research criteria is .47, in comparison with .36 for administrative criteria. This pattern of findings held for job-related interviews (.50 vs. .39) and job-related structured interviews (.51 vs. .37), but not for job-related unstructured interviews (.38 vs. .41). This contradiction to the general trend may be due to the relatively low number of coefficients in the distribution of job-related unstructured interviews with research criteria. Validities obtained based on criterion measures collected only for research purposes are typically larger than those based on criterion measures collected for administrative

purposes. Because the latter are more likely to contain biases and contaminants, we conclude that the validity estimates based on the research criteria are the more accurate and that those based on administrative criteria are substantially downwardly biased.

Summary, Conclusions, and Implications for Future Research

Validity may be concluded to be generalizable if the value at the lower 10th percentile of the distribution of estimated true validities is greater than zero (Callender & Osburn, 1981). This definition of validity generalizability is directly analogous to significance testing: A correlation is statistically significant when the lower bound of its confidence interval is above zero. By this criterion, almost all distributions (35 of 37) in Tables 4 through 8 show validity generalization.

However, some cautionary statements appear to be appropriate with respect to the unstructured interview. As noted earlier, the data summarized in this study come from employment interviews that are likely to be different from those normally conducted. The typical interview in the private sector is likely to be substantially less structured than the unstructured interviews used in this study because private sector interviews typically do not use scoring guides or result in quantifiable evaluations. Therefore, the validity estimates obtained in this study for unstructured interviews might overestimate the operational validity of many unstructured interviews used in business and industry. Those wishing to support their present interview practices on the basis of this study should review the cited primary studies to determine the extent to which their interview is similar to those analyzed in the present research.

Dreher, Ash, and Hancock (1988) took the opposite position. They concluded that most interview validity studies have underestimated the validity of interviews because researchers have usually failed to consider that individual interviewers differ in their ability to provide accurate predictions of employee behavior and in their tendencies to make favorable or unfavorable ratings. Dreher et al. argued that the procedure of collapsing data across multiple interviewers results in underestimates of the validity of interviews. To the extent that this effect is operating, it represents a downward bias in all of our validity estimates, including those for the unstructured interview. Despite this fact, it seems likely that the short, casual, conversational interviews one often encounters in business and industry will have mean validities lower than those reported in this study for the unstructured interview.

In this article, we have summarized the validity of various types of interviews. For job performance criteria, situational interviews yield the highest mean validity, followed by job-related and psychological interviews. However, for training performance, psychological interviews were similar in validity to (nonsituational) job-related interviews. The validity of the interview was found to vary by interview structure: Structured interviews yielded higher validities than did unstructured interviews. Validity was similar for job performance criteria and training performance criteria. Validity was lowest for tenure criteria, but inability to correct for criterion unreliability may explain this finding. Studies using criteria collected for research purposes generally yielded higher validities than those using administrative criteria, indicating that validity estimates computed with administrative criteria are underestimates.

Several findings are based on distributions containing few studies. Distributions with few coefficients have greater potential for second-order sampling error, which can distort true validity variance estimates and, to a lesser extent, can distort true mean validity estimates (Hunter & Schmidt, 1990, chap. 9; Schmidt et al., 1985, Question and Answer number 25). Therefore, these meta-analyses should be rerun in the future as more studies become available. This caveat, however, should not overly restrict conclusions drawn from our findings. Although some distributions contain fewer coefficients and smaller total sample sizes than desired, the impact of the interview and criterion characteristics is largely consistent across the various analyses. This consistency allows one to place more confidence in the findings.

This meta-analysis of employment interview validities is different from past meta-analyses of ability constructs (e.g., verbal ability) in two major ways. The first concerns the differentiation of constructs. The employment interview is a measurement method, as is a paper-and-pencil test. When one meta-analytically summarizes the validity of paper-and-pencil tests, one conducts the analyses separately for the different constructs measured by the tests. Separate meta-analyses are performed because the construct distinctions are considered meaningful in their own right and because different constructs may have different correlations with performance. Like paper-and-pencil tests, employment interviews may measure different constructs (e.g., cognitive ability, interpersonal skills, and manifest motivation). However, separate analyses for different constructs are not possible because of the dearth of employment interview validities reported for separate constructs.

The second difference between this study and other meta-analytic reviews is that there is more variability in how interview data are collected than there is in how ability data are collected. Although paper-and-pencil measures of a given ability may vary slightly (e.g., they may use different item types), the measurement process used for gathering data for a given ability is very similar. In contrast, employment interviews vary widely in data collection processes. Some interviewers follow procedures prescribed by their organization or by authors of how-to-interview publications, whereas others have no predetermined agenda. We have attempted to address these process distinctions by analyzing the data separately for structured and unstructured interviews. Although this distinction is a meaningful one for many interview developers and researchers (Asher, 1971; Hoeschen, 1977; Mayfield, 1964; McMurry, 1947; Tupes, 1950; Wagner, 1949; O. R. Wright, 1969), the resulting interview categories used in the present research are not perfectly homogeneous. For example, some structured interviews are more structured than others.

These differences between the present research and most past validity generalization studies suggest that the present research has less control over the constructs measured and the measurement process. These two sources of uncontrolled variance affect meta-analytic findings in that they increase the apparent situational specificity and reduce validity generalizability. Hence, conclusions based on such analyses tend to be conservative; they

overestimate situational specificity and underestimate validity generalizability.

The data and results presented in this article contrast somewhat with some traditional beliefs about the validity of the interview. In our experience, many personnel psychologists believe that the interview generally has low validity. This was our belief also, before we undertook this study. This belief was supported by the findings of Reilly and Chao (1982), Hunter and Hunter (1984), and Dunnette et al. (1971). Because our conclusions are contrary to much professional opinion and many previous quantitative reviews (except for Wiesner & Cronshaw, 1988, and P. M. Wright et al., 1989), our results should be examined carefully. The disparities between most past reviews and the present review can be explained, at least in part, by the effects of interview type and structure characteristics and by the effects of criterion content and purpose.

The meta-analysis findings for the validity of the interview reported here contrast in an interesting way with those reported for evaluations of education, training, and experience by McDaniel et al. (1988). In the case of both the interview and training and experience evaluations, the conventional belief has traditionally been that both had low (or zero) validity. Validity generalization analysis showed that this belief was essentially correct for the most commonly used method of training and experience evaluation: the point method, which had a mean true validity of only .13. However, in the case of the interview, the conclusion was the opposite: Even the unstructured interview was found to have a respectable level of validity.

Our results confirm conventional wisdom regarding the superiority of criterion measures collected for research purposes (Wherry & Bartlett, 1982). The finding that the administrative–research criterion dichotomy explains differences among studies in validity has implications for the interpretation of past validity generalization results and the conduct of future validity generalization research. This study has shown that validity coefficients based on research criteria are generally larger than those based on administrative criteria. Because this distinction has not been explicitly addressed in past validity generalization studies of cognitive abilities, the results of those studies are probably more conservative than was previously thought. The variance caused by this criterion characteristic can be expected to have caused an overestimate of the true standard deviations and an underestimate of the true operational validity of cognitive ability tests (Schmidt et al., 1993).

Although these analyses significantly advance knowledge of the validity of the interview, there is still a need for additional research. First, future validity studies should include a detailed description of the interview to permit a taxonomy of interview content and structure. Correlations between the interview scores and measures of constructs (e.g., cognitive ability) also should be included. Furthermore, future validity studies should report range-restriction and reliability information; at present this is done only sporadically. Because some types of interviews have been shown to yield validities as high as those for cognitive ability tests, we join Harris (1989) in calling for further study of the construct validity of the interview: To what extent does it measure motivation, social skills, and communication skills? At present, we know only that it measures multiple factors that predict job and training success. If these factors can be isolated, it may be possible to develop more reliable measurement methods for the underlying constructs (Hunter & Hunter, 1984). An important question for future researchers is the extent to which the interview can contribute incrementally over measures of cognitive ability, or vice versa.

References

Alexander, R. A., Carson, K. P., Alliger, G. M., & Cronshaw, S. F. (1989). Empirical distributions of range restricted SD_x in validity studies. *Journal of Applied Psychology, 74,* 253–258.

Anderson, R. C. (1954). The guided interview as an evaluative instrument. *Journal of Educational Research, 48,* 203–209.

Anstey, E. (1966). The civil service administrative class and the diplomatic service: A follow-up. *Occupational Psychology, 40,* 139–151.

Anstey, E. (1977). A 30-year follow-up of the CSSB procedure, with lessons for the future. *Journal of Occupational Psychology, 50,* 149–159.

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

Ash, R. A. (1981). Comparison of four approaches to the evaluation of job applicant training and work experience. *Dissertation Abstracts International, 42,* 4606B. (University Microfilms No. DE082-07909)

Asher, J. J. (1971). Reliability of a novel format for the selection interview. *Psychological Reports, 26,* 451–456.

Bartlett, N. R. (1950). Review of research and development in examination for aptitude for submarine training, 1942–1945. *Medical Research Laboratory Report, 153*(9), 11–53.

Bobbitt, J. M., & Newman, S. H. (1944). Psychological activities at the United States Coast Guard Academy. *Psychological Bulletin, 41,* 568–579.

Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Campbell, J. T., Otis, J. L., Liske, R. E., & Prien, E. P. (1962). Assessments of higher-level personnel: 2. Validity of the overall assessment process. *Personnel Psychology, 15,* 63–74.

Carlson, R. E. (1967). Selection interview decisions: The effect of interviewer experience, relative quota situation, and applicant sample on interviewer decisions. *Personnel Psychology, 20,* 259–280.

Carlson, R. E. (1970). Effect of applicant sample on ratings of valid information in an employment setting. *Journal of Applied Psychology, 54,* 217–222.

Castle, P. F. C., & Garforth, F. I., de la P. (1951). Selection, training and status of supervisors: 1. Selection. *Occupational Psychology, 25,* 109–123.

Dann, J. E., & Abrahams, N. M. (1970). *Use of biographical and interview information in predicting Naval Academy disenrollment.* (Research Rep. SRR 71-7). San Diego, CA: Naval Personnel and Training Research Laboratory.

Denton, J. C. (1964). The validation of interview-type data. *Personnel Psychology, 17,* 281–287.

Dicken, C. F., & Black, J. D. (1965). Predictive validity of psychometric evaluations of supervisors. *Journal of Applied Psychology, 49,* 34–47.

Dreher, G. F., Ash, R. A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the interview. *Personnel Psychology, 41,* 315–328.

Drucker, A. J. (1957). Predicting leadership ratings in the United States Army. *Educational and Psychological Measurement, 17,* 240–263.

Dunnette, M. D. (1970). *Multiple assessment procedures in identifying and developing managerial talent.* Minneapolis: University of Minnesota, Center for the Study of Organizational Performance and Human Effectiveness.

Dunnette, M. D., Arvey, R. D., & Arnold, J. A. (1971). *Validity study results for jobs relevant to the petroleum refining industry.* Minneapolis, MN: Personnel Decisions.

Dye, D. A. (1988). *What's in OPM's validity generalization data base?* (OPRD-88-7). Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development.

Felix, R. H., Cameron, D. C., Bobbitt, J. M., & Newman, S. H. (1945). An integrated medico-psychological program at the United States Coast Guard Academy. *American Journal of Psychiatry, 101,* 635–642.

Gardner, K. E., & Williams, A. P. O. (1973). A twenty-five year follow-up of an extended interview selection program in the Royal Navy. *Occupational Psychology, 47,* 1–13.

Glaser, R., Schwartz, P. A., & Flanagan, J. C. (1956). *Development of interview and performance tests for the selection of wage board supervisors* (Personnel Research Board Technical Research Note 53). Pittsburgh, PA: American Institutes for Research for Personnel Research Branch of the Adjutant General's Office.

Glaser, R., Schwartz, P. A., & Flanagan, J. C. (1958). The contribution of interview and situational performance procedures to the selection of supervisory personnel. *Journal of Applied Psychology, 42,* 69–73.

Grant, D. L., & Bray, D. W. (1969). Contributions of the interview to assessment of management potential. *Journal of Applied Psychology, 53,* 24–34.

Hakel, M. D., Ornesorge, J. P., & Dunnette, M. D. (1970). Interviewer evaluations of job applicants' resumes as a function of the qualification of the immediately preceding applicants. *Journal of Applied Psychology, 54,* 27–30.

Handyside, J. D., & Duncan, D. C. (1954). Four years later: A follow-up of an experiment in selecting supervisors. *Occupational Psychology, 28,* 9–23.

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

Hilton, A. C., Bolin, S. F., Parker, J. W., Jr., Taylor, E. K., & Walker, W. B. (1955). The validity of personnel assessments by professional psychologists. *Journal of Applied Psychology, 39,* 287–293.

Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39,* 399–420.

Hoeschen, P. L. (1977). *Using the Q by Q interview to predict success as a police sergeant.* Unpublished master's thesis, San Jose State University.

Huffcutt, A. I. (1992). *An empirical investigation of the relationship between multidimensional degree of structure and the validity of the employment interview.* Unpublished doctoral dissertation, Texas A&M University, College Station.

Hunt, W. A., Herrmann, R. S., & Noble, H. (1957). The specificity of the psychiatric interview. *Journal of Clinical Psychology, 13,* 49–53.

Hunter, J. E., & Hunter, R. F. (1984). The validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

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

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Newbury Park, CA: Sage.

Johnson, E. M. (1979). *The relationship between selection variables and performance variables in academy training for police officers.* Unpublished manuscript, City of Milwaukee, Fire and Police Commission, Wisconsin.

Johnson, E. M., & McDaniel, M. A. (1981, July). *Training correlates of a police selection system.* Paper presented at the International Personnel Management Association Assessment Council, Denver, CO.

Kennedy, R. (1986). *An investigation of criterion-related validity for the structured interview.* Unpublished master's thesis, East Carolina University, Greenville, NC.

Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice.* Newbury Park, CA: Sage.

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.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994). Non-linearity of range correlations in meta-analysis: A test of an improved procedure. *Journal of Applied Psychology, 79,* 425–438.

Ledvinka, J. (1973). Race of employment interviewer and reasons given by job seekers leaving their jobs. *Journal of Applied Psychology, 58,* 362–364.

Lilienthal, R. A., & Pearlman, K. (1983). *The validity of federal selection tests for aid/technicians in the health, science, and engineering fields.* Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development.

Matarazzo, R. G., Matarazzo, J. D., Saslow, G., & Phillips, J. S. (1958). Psychological test and organismic correlates of interview interaction patterns. *Journal of Abnormal and Social Psychology, 56,* 329–338.

Mayfield, E. C. (1964). The selection interview: A reevaluation of published research. *Personnel Psychology, 17,* 239–260.

McDaniel, M. A. (1986). Computer programs for calculating meta-analysis statistics. *Educational and Psychological Measurement, 46,* 175–177.

McDaniel, M. A., & Reck, M. (1985). *Validity generalization coding instructions* (Draft Report). Washington, DC: U.S. Office of Personnel Management, Office of Staffing Policy.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

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

Miles, D. W., Wilkins, W. L., Lester, D. W., & Hutchens, W. H. (1946). The efficiency of a high speed screening procedure in detecting the neuropsychiatrically unfit at a U.S. Marine Corps recruit training depot. *Journal of Psychology, 21,* 243–268.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., & Vaughn, M. J. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77,* 571–587.

Newmann, I., & Abrahams, N. M. (1979). *Validation of NROTC selection procedures* (NPRDC Special Rep. 79-12). San Diego, CA:Naval Personnel Research and Development Center.

Parrish, J. A., Klieger, W. A., & Drucker, A. J. (1955). *Assessment ratings and experimental tests as predictors of performance in officer candidate school* (Personnel Research Branch Technical Research Note 49). Washington, DC: Department of the Navy, Adjutant General's Office.

Pearlman, K. (1979). *The validity of tests used to select clerical personnel: A comprehensive summary and evaluation* (TS-79-1). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center. (NTIS No. PB 80.102650)

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict training success and job pro-

ficiency in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Pursell, E. D., Campion, M. A., & Gaylord, R. S. (1980). Structured interviewing: Avoiding selection problems. *Personnel Journal, 59,* 907–912.

Putney, R. W. (1947). Validity of the placement interview. *Personnel Journal, 23,* 144–145.

Reeb, M. (1968). Construction of a questionnaire to replace a valid structured interview in Israel Defence Forces. *Megamot: Behavioral Science Quarterly, 26,* 69–74.

Reeb, M. (1969). A structured interview for predicting military adjustment. *Occupational Psychology, 43,* 193–199.

Reilly, R. A., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–62.

Rimland, B. (1958). *The use of the interview in selecting NROTC students* (Project *SD* 1104.9.1). San Diego, CA: U.S. Naval Personnel Research Field Activity.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Rothstein, H. R., & McDaniel, M. A. (1989). Guidelines for conducting and reporting meta-analyses. *Psychological Reports, 65,* 759–770.

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: University of Maryland.

Ryan, A. M., & Sackett, P. R. (1989). Exploratory study of individual assessment practices: Interrater reliability and judgments of assessor effectiveness. *Journal of Applied Psychology, 74,* 568–579.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions and answers about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697–798.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, J. R., Pearlman, K., & McDaniel, M. A. (1993). Refinements in validity generalization procedures: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 3–13.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627–670.

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

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology, 23,* 251–263.

Taylor, E. L., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4,* 39–47.

Trankell, A. (1959). The psychologist as an instrument of prediction. *Journal of Applied Psychology, 43,* 170–175.

Tupes, E. C. (1950). An evaluation of personality-trait ratings obtained by unstructured assessment interviews. *Psychological Monographs, 64*(11, Whole No. 287).

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

Veres, J. G., III, Field, H. S., & Boyles, W. R. (1983). Administrative versus research performance ratings: An empirical test of rating data quality. *Public Personnel Management, 12,* 290–298.

Vernon, P. E. (1950). The validation of civil service selection board procedures. *Occupational Psychology, 24,* 75–95.

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 structure interview.* Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Warmke, D. L., & Billings, R. S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. *Journal of Applied Psychology, 64,* 124–131.

Warmke, D. L., & Weston, D. J. (1992). Success dispels myths about panel interviewing. *Personnel Journal, 71,* 120–126.

Weston, D. J., & Warmke, D. L. (1988). Dispelling the myths about panel interviews. *Personnel Administrator, 33,* 109–111.

Wexley, K. N., Yukl, G., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology, 56,* 45–48.

Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings. *Personnel Psychology, 35,* 521–551.

Whetzel, D. L., & McDaniel, M. A. (1988). Reliability of validity generalization databases. *Psychological Reports, 63,* 131–134.

Wiesner, W. H., & Cronshaw, S. F. (1988). The moderating impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61,* 275–290.

Wilson, N. A. B. (1948). The work of the civil service selection board. *Occupational Psychology, 22,* 204–212.

Wright, O. R. (1969). Summary of research on the selection interview since 1964. *Personnel Psychology, 22,* 391–413.

Wright, P. M., Lichtenfels, P. A., & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology, 62,* 191–199.

Zaccaria, M. A., Dailey, J. T., Tupes, E. C., Stafford, A. R., Lawrence, H. G., & Ailsworth, K. A. (1956). *Development of an interview procedure for USAF officer applicants* (AFPTRC-TN-56-43, Project No. 7701). Lackland Air Force Base, TX: Air Research and Development Command.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of appraisal. *Journal of Applied Psychology, 67,* 752–758.

Appendix

Sources of Validity Data

Albrecht, P. A., Glaser, G. M., & Marks, J. (1964). Validation of a multiple-assessment procedure for managerial personnel. *Journal of Applied Psychology, 48,* 351–360.

Aramburo, D. J. (1981). *The efficacy of the structured interview in the selection of special education teachers.* Unpublished doctoral dissertation, University of New Orleans, LA.

Ard, R. F. (1985, January 30). Personal communication to M. A. McDaniel.

Arvey, R. D., Miller, H. E., Gould, R., & Burch, P. (1987). Interview validity for selecting sales clerks. *Personnel Psychology, 40,* 1–12.

Banta, G. W. (1967). A comparison of the leaderless group discussion and the individual interview techniques for the selection of student orientation assistants. *Dissertation Abstracts International, 28,* 937.

Barrett, G. V., Svetlik, B., & Prien, E. P. (1966). Validity of the job-concept interview in an industrial setting. *Journal of Applied Psychology, 51,* 233–235.

Bartlett, N. R. (1950). Review of research and development in examination for aptitude for submarine training, 1942–1945. *Medical Research Laboratory Report 153*(9), 11–53.

Bender, W. R. G., & Loveless, H. E. (1958). Validation studies involving successive classes of trainee stenographers. *Personnel Psychology, 11,* 491–508.

Benz, M. P. (1974). *Validation of the examination for Staff Nurse II.* Urbana: University Civil Service Testing Program of Illinois, Testing Research Program.

Berkley, S. (1984). *VII Validation Report Corrections Officer Trainee.* Harrisburg: Commonwealth of Pennsylvania, State Civil Service Commission.

Bertram, F. D. (1975). *The prediction of police academy performance and on the job performance from police recruit screening measures.* Unpublished doctoral dissertation, Marquette University, Milwaukee, WI.

Bloom, R. F., & Brundage, E. G. (1947). Prediction of success in elementary schools for enlisted personnel. In D. B. Stuit (Ed.), *Personnel research and test development in the Bureau of Naval Personnel* (pp. 251–254). Princeton, NJ: Princeton University Press.

Bobbitt, J. M., & Newman, S. H. (1944). Psychological activities at the United States Coast Guard Academy. *Psychological Bulletin, 41,* 568–579.

Bolanovich, D. J. (1944). Selection of female engineering trainees. *Journal of Educational Psychology, 35,* 545–553.

Bonneau, L. R. (1957). An interview for selecting teachers (Doctoral dissertation, University of Nebraska, 1956). *Dissertation Abstracts International, 17,* 537–538.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67,* 3–9.

Bosshardt, M. J. (1992). *Situational interviews versus behavior description interviews: A comparative validity study.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Bosshardt, M. J. (1993, April 8). Personal communication to M. A. McDaniel.

Brown, S. H. (1986). Personal communication to M. A. McDaniel.

Campbell, J. T., Prien, E. P., & Brailey, L. G. (1960). Predicting performance evaluations. *Personnel Psychology, 13,* 435–440.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41,* 25–42.

Carlstrom, A. (1984). *Correlations between selection ratings and success as army aviators* (FOA Rep. C 53020-H2). Linkoping, Sweden: National Defense Research Institute.

Cerf, A. Z. (1947). Conference for the interpretation of test scores and occupational background CE707A. In J. P. Guilford & J. I. Lacey (Eds.), *Printed classification tests: Report No. 5* (pp. 652–656). Washington, DC: Army Air Forces Aviation Psychology Program.

Conrad, H. S., & Satter, G. A. (1946). *The use of test scores and quality-classification ratings in predicting success in electrician's mates school* (Office of Scientific Research and Development Rep. No. 5667, 1945: Publications Board No. 13290). Washington, DC: U.S. Department of Commerce.

Cook, R. E. (1981). *Chippewa Valley local validation study of the teacher perceiver interview process.* Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Darany, T. (1971). *Summary of state police trooper 07 validity study.* Ann Arbor: Michigan Department of Civil Service.

Davey, B. (1984, May). *Are all oral panels created equal?: A study of differential validity across oral panels.* Paper presented at the International Personnel Management Association Assessment Council Annual Conference, Seattle, WA.

Davis, R. (1986). Personal communication to M. A. McDaniel.

DeHart, W. A. (n.d.). *Preliminary report on research findings of a demonstration project in recruitment and training of case workers sponsored by the Utah State Department of Public Welfare, 1966—1967.* Salt Lake City: Utah State Department of Public Welfare.

Delaney, E. C. (1954). Teacher selection and evaluation with special attention to the validity of the personal interview and the National Teacher Examinations as used in one selected community (Elizabeth, New Jersey). (Doctoral dissertation, Columbia University, 1954). *Dissertation Abstracts International, 14,* 1334–1335.

Delery, J. E., Wright, P. M., & Tolzman, K. (1992, April). *Employment tests and the situational interview: A test of incremental validity.* Paper presented at the seventh annual meeting of the Society for Industrial and Organizational Psychology, Montreal, Quebec, Canada.

Dipboye, R. L., Gaugler, B., & Hayes, T. (1990, April). *Differences among interviewers in the incremental validity of their judgments.* Paper presented at the fifth annual meeting of the Society for Industrial and Organizational Psychology, Miami, FL.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9–15.

Dubois, P. H., & Watson, R. K. (1950). The selection of patrolman. *Journal of Applied Psychology, 34,* 90–95.

Dunlap, J. W., & Wantman, M. J. (1944). *An investigation of the interview as a technique for selecting aircraft pilots* (Rep. No. 33). Washington, DC: Civil Aeronautics Administration, Airman Development Division.

English, J. J. (1983). *The relationship between structured selection criteria and the assessment of proficient teaching.* Unpublished doctoral dissertation, University of Virginia, Richmond.

Finesinger, J. E., Cobb, S., Chapple, E. D., & Brazier, M. A. B. (1948). *An investigation of prediction of success in naval flight training* (Rep. No. 81). Washington, DC: Civil Aeronautics Administration, Division of Research.

Fisher, J., Epstein, L. J., & Harris, M. R. (1967). Validity of the psychiatric interview: Predicting the effectiveness of the first Peace Corps volunteers in Ghana. *Archives of General Psychiatry, 17,* 744–750.

Flynn, J. T., & Peterson, M. (1972). The use of regression analysis in police patrolman selection. *The Journal of Criminal Law, Criminology, and Police Science, 63,* 564–569.

(Appendix continues on next page)

Fowlkes, R. D. (1984). *The relationship between the teacher perceiver interview and instructional behaviors of teachers of learning disabled students.* Unpublished doctoral dissertation. University of Virginia, Richmond.

Freeman, G. L., Manson, G. E., Katzoff, E. T., & Pathman, J. H. (1942). The stress interview. *Journal of Abnormal and Social Psychology, 37,* 427–447.

Friedland, D. (1973). *Selection of police officers.* Los Angeles: City of Los Angeles, Personnel Department.

Friedland, D. (1980). *A predictive study of selected tests for police officer selection.* Los Angeles: City of Los Angeles, Personnel Department.

Friedland, D. (n.d.). *Section II—City of Los Angeles Personnel Department (Junior Administrative Assistant Validation Study).* Los Angeles: City of Los Angeles, Personnel Department.

Ghiselli, E. E. (1966). The validity of a personnel interview. *Personnel Psychology, 19,* 389–394.

Gillies, T. K. (1988). *The relationship between selection variables and subsequent performance ratings for teachers in an Oregon school district.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Githens, W. H., & Rimland, B. (1964). *The validity of NROTC selection interviews against career decisions and officer fitness reports: An eight year followup* (PRASD Rep. No. 234). San Diego, CA: U.S. Naval Personnel Research Activity.

Glaser, R., Schwartz, P. A., & Flanagan, J. C. (1958). The contribution of interview and situational performance procedures to the selection of supervisory personnel. *Journal of Applied Psychology, 42,* 69–73.

Green, P. C., & Alter, P. (1990). *Validity of a multi-position, Behavioral Interviewing system when using an answer scoring strategy.* Memphis, TN: Behavioral Technology.

Gregory, E. (1956). *Evaluation of selection procedures for women naval officers* (USN Bureau of Naval Personnel Technical Bulletin, No. 56-11). [Cited in Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 55,* 178–200.]

Grove, D. A. (1981). A behavioral consistency approach to decision making in employment selection. *Personnel Psychology, 34,* 55–64.

Halliwell, S. T. (1984, November). *An evaluation of the Canadian Forces combat arms officer selection board.* Paper presented at the 26th annual conference of the Military Testing Association, Munich, Germany (Available from Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada)

Harris, J. G. (1972). Prediction of success on a distant pacific island: Peace Corps style. *Journal of Consulting and Clinical Psychology, 38,* 181–190.

Hays, E. J. (1990). *Relationship of a situational interview to the job performance of convenience store clerks.* Unpublished master's thesis, University of North Texas, Denton.

Hess, L. R. (1973). Police entry tests and their predictability of score in police academy and subsequent job performance (Doctoral dissertation, Marquette University, 1972). *Dissertation Abstracts International, 24,* 679–686.

Hill, C. J., Jr., Russell, D. L., & Wendt, G. R. (1946). *A partial analysis of a three-man interview technique for predicting airline pilot success in the civilian pilot training program* (Rep. No. 65). Washington, DC: Civil Aeronautics Administration, Division of Research.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 53,* 1–12.

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

Huse, E. F. (1962). Assessments of higher-level personnel: IV. The validity of assessment techniques based on systematically varied information. *Personnel Psychology, 15,* 195–205.

Husen, T. (1954). La validité des interviews par rapport à l'age, au sexe et à la formation des interview. *Le Travail Humain, 17,* 60–67.

Inman, S. J. (1986). Personal communication to M. A. McDaniel regarding the study by the city of Milwaukee, WI: The selection and evaluation of unskilled laborers.

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

Johnson, E. M., & McDaniel, M. A. (1981, June). *Training correlates of a police selection system.* Paper presented at the Annual Conference of the International Personnel Management Council, Denver, CO.

Johnson, G. (1990). *The structured interview: Manipulating structuring criteria and the effects on validity, reliability, and practicality.* Unpublished doctoral dissertation, Tulane University, New Orleans, LA.

Johnson, H. M., Boots, M. L., Wherry, R. J., Hotaling, O. C., Martin, L. G., & Cassens, F. P., Jr. (1944). *On the actual and potential value of biographical information as a means of predicting success in aeronautical training* (Rep. No. 32). Washington, DC: Civil Aeronautics Administration, Airman Development Division.

Jones, D. E. (1978). *Predicting teaching processes with the teacher perceiver interview.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg.

Kelley, E. L., & Ewart, E. (1942). *A preliminary study of certain predictors of success in civilian pilot training* (Rep. No. 7). Washington, DC: Civil Aeronautics Administration, Division of Research.

Kelly, E. L., & Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist, 5,* 395–406.

Kennedy, R. (1986). *An investigation of criterion-related validity for the structured interview.* Unpublished master's thesis, East Carolina University, Greenville, NC.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment: Fairness and validity of personnel tests for different ethnic groups.* New York: New York University Press.

Knights, R. M. (1977). The relationship between the selection process and on-the-job performance of Albuquerque police officers (Doctoral dissertation, University of New Mexico, 1976). *Dissertation Abstracts International, 38,* 1229A-1230A.

Komives, E., Weiss, S. T., & Rosa, R. M. (1984). The applicant interview as a predictor of resident performance. *Journal of Medical Education, 59,* 425–426.

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

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,* 422–426.

Lopez, F. M. (1966). Current problems in test performance of job applicants: I. *Personnel Psychology, 19,* 10–18.

Martin, M. A. (1972). *Reliability and validity of Canadian Forces selection interview procedures* (Rep. 72-4). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Maurer, S. (1986). Personal communication to M. A. McDaniel.

McKinney, T. S. (1975). *The criterion-related validity of entry level police officer selection procedures.* Phoenix, AZ: City of Phoenix, Personnel Department, Employee Selection Research.

McKinney, T. S., & Meltzzer, D. M. (1979). *The validity and utility of entry level police recruit selection procedures: The Phoenix Experience.* Phoenix, AZ: City of Phoenix, Personnel Department, Employee Selection Research.

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

Merenda, P. F., & Clarke, W. V. (1959). AVA validity for textile workers. *Journal of Applied Psychology, 43,* 162–165.

Meyer, H. H. (1956). An evaluation of a supervisory selection program. *Personnel Psychology, 9,* 499–513.

Miner, J. B. (1970). Psychological evaluations as predictors of consulting success. *Personnel Psychology, 23,* 393–405.

Mischel, W. (1965). Predicting the success of Peace Corps volunteers in Nigeria. *Journal of Personality and Social Psychology, 1,* 510–517.

Modderaar, J. (1966). Over de validiteit van de selectie voor bedrijven [About the validity of selection for industry]. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 21,* 573–589.

Mosher, M. R. (1991, June). *Development of a behaviorly consistent structured interview.* Paper presented at the 27th International Applied Military Psychology Symposium, Stockholm, Sweden.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., & Vaughn, M. J. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77,* 571–587.

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

Parrish, J. A., Klieger, W. A., & Drucker, A. J. (1955). *Assessment ratings and experimental tests as predictors of performance in officer candidate school* (Personnel Research Branch Technical Research Note 49). Washington, DC: Department of the Army, Adjutant General's Office.

Pasco, D. C. (1979). The comparison of two interviewing techniques, the leaderless group discussion and the traditional interview, as a method for teacher selection (Doctoral dissertation, University of Maryland, 1979). *Dissertation Abstracts International, 41,* 486.

Pashalian, S., & Crissy, W. J. E. (1953). *The interview IV: The reliability and validity of the assessment interview as a screening and selection technique in the submarine service* (Medical Research Laboratory Rep. No. 216, XII, No. 1). Washington, DC: Department of the Navy, Bureau of Medicine and Surgery.

Pearlman, K. (1978). Personal communication to M. A. McDaniel.

Plag, J. A. (1961). Some considerations of the value of the psychiatric screening interview. *Journal of Clinical Psychology, 17,* 3–8.

Prien, E. P. (1962). Assessments of higher-level personnel: V. An analysis of interviewers' predictions of job performance. *Personnel Psychology, 15,* 319–334.

Pulos, L., Nichols, R. C., Lewinsohn, P. M., & Koldjeski, T. (1962). Selection of psychiatric aides and prediction of performance through psychological testing and interviews. *Psychological Reports, 10,* 519–520.

Rafferty, J. A., & Deemer, Jr., W. L. (1950). Factor analysis of psychiatric impressions. *Journal of Educational Psychology, 41,* 173–183.

Raines, G. N., & Rohrer, J. H. (1955). The operational matrix of psychiatric practice: Consistency and variability in interview impressions of different psychiatrists. *The American Journal of Psychiatry, 3,* 721–733.

Reeb, M. (1969). A structured interview for predicting military adjustment. *Occupational Psychology, 43,* 193–199.

Requate, F. (1971–1972). *Metropolitan Dade County fire fighter validation study, Miami, Florida.* Miami: Dade County Personnel Department.

Rhea, B. D. (1966). *Validation of OCS selection instruments: The relationship of OCS selection measures to OCS performance* (Tech. Bulletin STB 66-18). San Diego, CA: U.S. Naval Personnel Research.

Rhea, B. D., Rimland, B., & Githens, W. H. (1965). *The development and evaluation of a forced-choice letter of reference form for selecting officer candidates* (Tech. Bulletin STB 66-10). San Diego: U.S. Naval Personnel Research Activity.

Ricchio, L. K. (1980). Evaluation of eligible list ranking as a predictor of teacher success. (Doctoral dissertation, Claremont Graduate School, 1980). *Dissertation Abstracts International, 41,* 50.

Robertson, I. T., Gratton, L., & Rout, U. (1990). The validity of situational interviews for administrative jobs. *Journal of Organizational Behavior, 11,* 69–76.

Roth, P. L., Campion, J. E., & Francis, D. J. (1988, April). *A LISREL analysis of the predictive power of pre-employment tests and the panel interview.* Paper presented at the third annual meeting of the Society for Industrial and Organizational Psychology, Dallas, Texas. (Validity data not reported in the paper were obtained from P. L. Roth.)

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: University of Maryland.

Shaw, J. (1952). The function of the interview for determining fitness for teacher-training. *Journal of Educational Research, 45,* 667–681.

Simmons, J. E. (1976). *A study to test teacher perceiver interview as an instrument that would select vocational agriculture instructors that develop positive rapport with their students.* Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Smedberg, C. A. (1984). *An examination of the validity of the empathy interview as an instrument for teacher selection.* Unpublished doctoral dissertation, University of South Florida, Tampa.

Solomon, G. L. (1981). *The Omaha teacher pre-employment interview as a predictor of job performance of selected new teachers in the Jefferson County school system.* Unpublished doctoral dissertation, University of Alabama, Tuscaloosa.

Sparks, C. P. (n.d.). [Title unknown]. Cited in Dunnette, M. D., Arvey, R. D., & Arnold, J. A. (1971). Validity study results for jobs relevant to the petroleum industry. Minneapolis, MN: Personnel Decisions.

Stephen, M. J. (1980). The selection of computer salespersons: An evaluation of a structured interview and its underlying theory. (Doctoral dissertation, Northwestern University, 1980). *Dissertation Abstracts International, 41,* 2372.

Stohr-Gillmore, M. K., Stohr-Gillmore, M. W., & Kistler, N. (1990). Improving selection outcomes with the use of situational interviews: Empirical evidence from a study of correctional officers for new generation jails. *Review of Public Personnel Administration, 10,* 1–18.

Tarico, V. S., Altmaier, E. M., Smith, W. L., Franken, E. A., & Berbaum, K. S. (1986). Development and validation of an accomplishment interview for radiology residents. *Journal of Medical Education, 61,* 845–847.

Trankell, A. (1959). The psychologist as an instrument of prediction. *Journal of Applied Psychology, 43,* 170–175.

Trites, D. K. (1960). Adaptability measures as predictors of performance ratings. *Journal of Applied Psychology, 44,* 349–353.

Tubiana, J. H., & Ben-Sitakhar, G. (1982). An objective group questionnaire as a substitute for a personal interview in prediction of success in military training in Israel. *Personnel Psychology, 35,* 349–357.

Tziner, A., & Dolan, S. (1982a). Evaluation of a traditional selection system in predicting success of females in officer training. *Journal of Occupational Psychology, 55,* 269–275.

Tziner, A., & Dolan, S. (1982b). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67,* 728–736.

U.S. Office of Personnel Management (1987). *The structured interview.* Washington, DC: Office of Examination Development, Alternative Examining Procedures Division.

Waldron, L. A. (1974). The validity of an employment interview independent of psychometric variables. *Australian Psychologist, 9,* 68–77.

Walters, L. C., Miller, M., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *Internal Journal of Aviation Psychology, 3,* 25–38.

Wayne County Civil Service Commission. (1973). *Summary informa-*

(Appendix continues on next page)

tion regarding validity of the examination for patrolmen in the Wayne County Sheriff's Department. Wayne County Civil Service Commission.

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72,* 484–487.

Willkins, W. L., Anderhalter, O. F., Rigby, M. K., & Stinson, P. (1955). *Statistical description of criterion measures for USMC junior officers* (Office of Naval Research Contract N7ONR-40802; NR 151-092). St. Louis, MO: St. Louis University, Department of Psychology.

Woodworth, D. G., Barron, F., & MacKinnon, D. W. (1957). *Analysis of life history interview's rating for 100 air force captains* (Research Rep. AFPTRC-TN, ASTIA Document AD 146-401). Lackland Air Force Base, TX: Air Research and Development Command.

Wright, P. M., Pursell, E. D., Lichtenfels, P. A., & Kennedy, R. (1986, August). *The structured interview: Additional studies and a meta-analysis.* Paper presented at the Academy of Management Convention, Chicago.

Wunder, S. (1973a). *Validity of a selection procedure for process trainees.* Unpublished manuscript.

Wunder, S. (1973b). *Validity of a selection program for refinery technician trainees.* Unpublished manuscript.

Wunder, S. (1974). *Validity of a selection program for commercial drivers.* Unpublished manuscript.

Wunder, S. (1977a). *The refinery and chemical plant test battery: Validity for laboratory selection.* Unpublished manuscript.

Wunder, S. (1977b). *Selecting process operator and laboratory trainees.* Unpublished manuscript.

Wunder, S. (1978a). *Preemployment tests for pipefitter trainees.* Unpublished manuscript.

Wunder, S. (1978b). *Selecting laboratory technicians.* Unpublished manuscript.

Wunder, S. (1979). *Preemployment tests for electrician trainees.* Unpublished manuscript.

Wunder, S. (1980). *Validity of preemployment measures for instrument technician trainees.* Unpublished manuscript.

Wunder, S. (1981). *Selection of manufacturing technicians.* Unpublished manuscript.

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

Zaranek, R. J. (1983). *A correlational analysis between the teacher perceiver interview and teacher success in the Chippewa Valley school system.* Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Zedeck, S., Tziner, A., & Middlestadt, S. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology, 36,* 355–370.

Received October 7, 1991
Revision received November 29, 1993
Accepted November 30, 1993 ■