

Running head: STRUCTURED INTERVIEWING

A Review of Structure in the Selection Interview

Michael A. Campion and David K. Palmer

Purdue University

and

James E. Campion

University of Houston

Abstract

Virtually every previous review has concluded that structuring the selection interview improves its psychometric properties. This paper reviews the research literature in order to describe and evaluate the many ways interviews can be structured. Based on nearly 200 articles and books, 15 components of structure are identified that may enhance either the content of the interview or the evaluation process in the interview. Each component is explained in terms of its various operationalizations in the literature. Then, each component is critiqued in terms of its impact on numerous forms of reliability, validity, and user reactions. Finally, recommendations for research and practice are presented. It is concluded that interviews can be easily enhanced by using some of the many possible components of structure, and the improvement of this popular selection procedure should be a high priority for future research and practice.

A Review of Structure in the Selection Interview

In the 80-year history of published research on employment interviewing (dating back to Scott, 1915), few conclusions have been more widely supported than the idea that structuring the interview enhances reliability and validity.

Brief Summary of Previous Reviews

Early narrative reviews were particularly graphic in their descriptions of the benefits of structure. Wagner (1949) stated that all interviews should be conducted according to a "standardized form" because "this prevents aimless rambling, lengthy digressions, and the possibility of omitting important areas" (p. 42). Mayfield (1964) concluded that "in almost all cases where a satisfactory reliability for the selection interview was reported, the interview was of a structured form" (p. 250). Finally, in recommending structured interviews, Ulrich and Trumbo (1965) noted that "...it is difficult to see how the interviewer can arrive at anything like an optimal strategy if the information available to him is continually different in kind" (p. 112). Subsequent narrative reviews have continued to support the use of structure (Arvey & J. Campion, 1982; Harris, 1989; Schmitt, 1976; Wright, 1969)

Meta-analytic reviews of validity studies have also unanimously supported the superiority of structured interviews. They differed somewhat in the studies they summarized and in the corrections they used for range restriction and unreliability, but their overall findings were very similar. Wiesner and Cronshaw (1988) analyzed 87 validity coefficients and found validities of .34 (.62 corrected) for structured interviews and .17 (.31) for unstructured interviews. Wright, Lichtenfels, and Pursell (1989) reviewed 13 coefficients for structured interviews and found a validity of .27 (.35 corrected for unreliability only), which they compared to an estimate of .14 for unstructured interviews (J. Hunter & R. Hunter, 1984). Huffcutt and Arthur (1994) summarized 114 coefficients in relation to degree of structure and found that validity ranged from .11 (.20 corrected) for the lowest level to .34 (.57) for the highest level of structure. McDaniel, Whetzel, Schmidt, and Maurer (1994) summarized 145 coefficients and found a validity of .24 (.44 corrected) for

structured compared to .18 (.33) for unstructured interviews. Marchese and Muchinsky (1993) summarized 31 coefficients and showed that validity was correlated .45 with degree of structure. Finally, Conway, Jako, and Goodman (1995) summarized 160 reliability coefficients and showed that reliability was correlated from .26 to .56 with degree of structure. They also estimated that reliability placed an upper limit on validity of .67 for highly structured and .34 for unstructured interviews.

Overview of Paper

Conceptualizations of structure are fragmented in the literature, with different researchers using different approaches. Previous reviews have not thoroughly examined the various ways interviews can be structured. The purpose of this paper is to intensively review the literature in order to summarize, integrate, and evaluate the many ways interviews can be structured to improve psychometric properties.

This paper complements previous narrative reviews in that it is not an inclusive examination of all research topics since the last review; instead, it focuses only on structure but considers the entire literature. The review by Dipboye and Gaugler (1993) similarly examined structure, but the present paper differs by focusing on psychometric (as opposed to behavioral and cognitive) consequences and by considering a broader range of structural components.

This paper complements meta-analytic reviews by offering potential explanations for higher validities. Also, meta-analyses have used only very general distinctions in terms of structure and have not enhanced our conceptual understanding. Furthermore, many components of structure have received too little empirical attention to allow meta-analyses, and those that have been studied tend to be confounded, thus not allowing meta-analytic tests.

This paper attempts to meet the criteria for literature reviews (M. Campion, 1993). First, it summarizes a large body of literature in a thorough and inclusive fashion. Second, it goes beyond previous reviews because of its unique focus. Third, it analyzes the literature critically as needed. Fourth, it organizes and explains findings across studies. Fifth, it identifies

limitations in the literature and defines questions for future research. Finally, it casts the literature within a single framework consisting of components of structure and their psychometric implications.

The literature for this review was drawn from the citations in previous reviews, computerized and manual searches, and cross-referencing. Nearly 200 articles and books that address the topic are examined in this paper.

The paper will use the term "structured" interviews, but other terms have been used such as "standardized," "guided," "systematic," and "patterned." The paper defines "structure" very broadly as any enhancement of the interview that increases psychometric properties by assisting the interviewer in determining what questions to ask or how to evaluate responses. The review of the literature yielded 15 components of structure that are listed in Table 1 and form the framework of the paper. The purpose is to objectively describe and evaluate these components of structure, but not to advocate that they all be used in every situation.

The 15 components are divided into two categories. The first includes components that influence the content of the interview. They structure the nature of the information that is elicited during the interview. The second category includes components that influence the evaluation process. They help the interviewer judge the information that is elicited. Although some components could be included in both categories, this distinction is useful because it highlights a primary difference between the components.

The impact of each component is evaluated in terms of reliability, validity, and user reactions. Eight types of reliability are considered:

1. Test-retest reliability of the content of the interview -- Is the same interview content elicited each time by the interviewer?
2. Test-retest reliability of the evaluation of the interview -- Is the same evaluation process consistently used each time by the interviewer?
3. Interrater reliability of the content of the interview -- Do different interviewers elicit the same content?
4. Interrater reliability of the evaluation of the interview -- Do

different interviewers evaluate candidates consistently?

5. Candidate consistency -- Does the interview elicit consistent responding from the candidate? Would transient variability in the candidate's interviewing skills, mood, stress, or similar factors influence the results?

6. Interviewer-candidate interaction -- Does the interview limit possible error variance due to differences in interactions between interviewers and candidates? Would differences or similarities between interviewers and candidates in personalities, communication styles, interpersonal attraction, or similar factors influence the results? Variation may be possible in the emergent relationship between interviewers and candidates (Dipboye, 1992).

7. Internal consistency -- Are the interview items sufficiently numerous and intercorrelated such that the composite measures a homogeneous construct?

8. Interrater agreement -- Do interviewers agree on their ratings or judgments? Is the difference in level between their ratings small, such that similar decisions would be made? Reliability refers to covariation, and is thus not the same as agreement which refers to mean differences (Tinsley & Weiss, 1975). Components that enhance reliability may not enhance agreement.

Three types of validity information are considered:

1. Job-relatedness -- Is the interview related to the content of the job?
2. Reduced deficiency -- Is measurement deficiency reduced? Does the interview elicit a large amount of useful information?
3. Reduced contamination -- Does the interview prevent contamination (e.g., faking or irrelevant information) from entering the process?

Three types of user reactions are also considered:

1. Reduced EEO bias -- Will the components of structure reduce potential bias against subgroups of candidates protected by equal employment opportunity (EEO) laws? This includes reducing potential adverse impact and disparate treatment, as well as increasing perceptions of fairness. Some structured interviews have been specifically developed to enhance legal defensibility (Pursell, M. Campion, & Gaylord, 1980). Research has found many components of structure to be empirically related to court verdicts (J. Campion & Arvey,

1989; Gollub-Williamson, J. Campion, Malos, M. Campion, & Roehling 1995).

2. Candidate reactions -- Will candidates view the interview positively? There is increasing recognition of the importance of candidate reactions to selection procedures (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Such reactions reflect the perceived (face) validity of selection procedures, and they influence job choice, affective reactions, and referrals of other candidates. Candidates may prefer interviews over some psychological tests (Wilson, 1948), but structured interviews may not elicit as positive a response as unstructured interviews (Latham & Finnegan, 1993).

3. Interviewer reactions -- Will interviewers view the interview positively? These include reactions to face validity and usability. Managers may recognize the job-relatedness and defensibility of structured interviews (Latham & Finnegan, 1993), but there may be political forces that encourage them to use unstructured interviews (Dipboye, 1994).

The 15 components of structure are reviewed individually below. First, the component is explained, and alternative operationalizations reflecting different levels of structure are described. The literature is exhaustively summarized in terms of how it illustrates the levels of structure. Second, the component is examined and critiqued on each type of reliability, validity, and user reaction (as shown in Table 2). Third, research and practice issues and recommendations are presented.

Review of Components of Structure

1. Base Questions on a Job Analysis

Explanation and Alternatives. A variety of job analysis methods have been used to develop structured interviews. Both worker- and job-oriented methods have value (McCormick, 1976). Worker-oriented methods determine the knowledge, skills, abilities, and other attributes upon which to develop interview questions. Job-oriented methods determine tasks, equipment, and situations that allow the questions to be worded in the proper terminology and context.

Critical incidents job analysis (Flanagan, 1954) is the most common approach mentioned in interviewing articles (M. Campion, J. Campion, & Hudson,

1994; M. Campion, Pursell, & Brown, 1988; Delery, Wright, McArthur, & Anderson, 1994; Janz, 1982; Latham & Saari, 1984; Latham, Saari, Pursell, & M. Campion, 1980; Latham & Skarlicki, 1995; Motowidlo, Carter, Dunnette, Tippins, Werner, Burnett, & Vaughan, 1992; Robertson, Gratton, & Rout, 1990; Schmitt & Ostroff, 1986; Weekley & Gier, 1987; Zedeck, Tziner, & Middlestadt, 1983). Typically, meetings with job experts were used to collect incidents, but surveys were also used (Weekley & Gier, 1987).

The value of critical incidents is that they provide ideas for interesting and job-related interview questions. However, the development of questions from incidents is part of the art or, at least, the unwritten aspects of structured interviewing. Some authors acknowledge that "literary license" is needed (Latham & Saari, 1984, p. 569). Incidents are often grouped into dimensions first (Motowidlo et al., 1992; Robertson et al., 1990), then the incidents that best represent the dimensions are turned into questions (Latham et al., 1980). This presumably enhances content validity.

Analogous to critical incidents analysis, contrasting groups of high and low performing employees are sometimes examined as an approach to job analysis (Holt, 1958; Robertson et al., 1990). One such technique is repertory grid analysis (V. Stewart & A. Stewart, 1981). It asks job experts to consider employees in groups of three and try to identify aspects of performance-related behavior that are similar between two employees and different from the third (Robertson et al., 1990). Others suggest that a behavioral consistency approach (Wernimont & Campbell, 1968) is a useful way of developing interview questions from job analyses (Feild & Gatewood, 1989; Grove, 1981; Schmitt & Ostroff, 1986).

Finally, there is the issue of who writes the questions. Some articles state that job experts are used, such as incumbents, managers, and interviewers (Janz, 1982; Latham et al., 1980; Orpen, 1985; Roth & J. Campion, 1992). Most articles do not discuss it, thus implying they are written by researchers.

There are at least three (unstructured) alternatives to job analysis. First, many interviews are conducted by psychologists and focus on personality

traits but are not based on job analysis (Bobbitt & Newman, 1944; Fisher, Epstein, & M. Harris, 1967; J. Harris, 1972; Hilton, Bolin, Parker, Taylor, & Walker, 1955; Mischel, 1965; Plag, 1961; Raines & Rohrer, 1955; Waldron, 1974; but cf. Holt, 1958). Second, interviewers may ask traditional questions that are common in unstructured interviews but not based on job analysis (e.g., Tell me about yourself? What are your strengths/weaknesses? What are your goals?). Third, an intuitive approach can be used wherein interviewers ask whatever questions are thought relevant.

Effects on Reliability, Validity, and User Reactions. Job analysis is a basic requirement for developing valid selection procedures according to both professional (Society for Industrial and Organizational Psychology, 1987) and legal (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) testing guidelines. Its value for structuring interviews was recognized very early (McMurry, 1947), yet it is noteworthy that relatively few articles mention it.

Job analysis is not expected to enhance reliability, even though there might be a weak positive relationship if it limits the domain of inquiry in the interview. The Conway et al. (1995) meta-analysis showed a low positive relationship between job analysis and reliability, which they interpreted as suggesting an indirect effect. It is expected to influence all three types of validity, however (Table 2). By definition, job analysis should enhance job-relatedness. This may occur partly because it allows the interviewer to obtain job-related samples of applicant behavior (Dipboye & Gaugler, 1993). A job analysis should enhance the amount of job information brought into the interview, thus decreasing deficiency. Similarly, by focusing the interview on job-related content, it should reduce contamination. Without a job analysis to provide interviewers with a common frame of reference, they might base the interview on idiosyncratic beliefs about job requirements (Dipboye, 1994).

The three most recent narrative reviews have emphasized the importance of job analysis for improving validity (Arvey & J. Campion, 1982; Harris, 1989; Schmitt, 1976). Likewise, the meta-analytic review of Wiesner and Cronshaw

(1988) found an average uncorrected validity of .48 when a formal job analysis was conducted, .35 when an informal analysis was conducted, and .31 when it was unknown if an analysis was conducted. McDaniel et al. (1994) found average uncorrected validities of .21 and .27 for interviews based on a job analysis and .15 for psychological interviews not based on a job analysis. There is also evidence that interview questions with higher content validity may have higher criterion-related validity (Carrier, Dalessio, & Brown, 1990).

Job analysis is expected to enhance all user reactions (Table 2). It has been shown to reduce EEO bias (Kesselman & Lopez, 1979) and help defend organizations in court (Kleiman & Faley, 1985) for other selection procedures. It is likely to do the same for interviews (Arvey & Faley, 1988; J. Campion & Arvey, 1989), and there is evidence to this effect from court cases on interviewing (Gollub-Williamson et al., 1995). Job-relatedness should enhance perceptions of face validity for both candidates and interviewers. Finally, involving interviewers in the job analyses should improve their acceptance.

Research and Practice Issues. Job analysis determines the content of the interview. If job analysis is more likely to identify knowledges, skills, and abilities, rather than personality traits and other attributes (Harvey, 1991), then a key question is raised: Are structured interviews more valid than unstructured interviews because they tap into cognitive ability? This is important because tests of cognitive ability are inexpensive and available.

Structured interviews have shown both high and low correlations with cognitive ability tests. M. Campion et al. (1988; 1994) found correlations of .43 and .60. Conversely, Pulakos and Schmitt (1995) found a correlation of .09, and Motowidlo et al. (1992) found a correlation with grades and class rank of .15. These results are not due to the apparent constructs assessed by the interviews. M. Campion et al. (1994) were attempting to measure attributes not usually considered cognitive abilities (e.g., teamwork, commitment, safety orientation, etc.), while Pulakos and Schmitt were measuring attributes that appeared cognitive (e.g., planning, problem solving, communicating, etc.). The studies differed in other ways that offer more plausible explanations for these

differences. In particular, the M. Campion et al. studies used more highly structured interviews and samples with a wider range of cognitive ability.

A recent meta-analysis of 48 studies found a correlation of .34 (.40 corrected) between interview ratings and ability test scores (Huffcutt, Roth, & McDaniel, 1995). Ratings with higher cognitive loading were also more valid.

Another approach to this issue is to examine whether structured interviews have incremental validity beyond cognitive ability tests. Again, the results have been equivocal, with some studies finding incremental validity (M. Campion et al., 1994; Pulakos & Schmitt, 1995) and others not (M. Campion et al., 1988; Delery et al., 1994; Walters, Miller, & Ree, 1993). The findings are not due simply to the cognitive loading of the interview; M. Campion et al. (1994) had the highest correlation with cognitive tests and Pulakos and Schmitt had the lowest, and both showed incremental validity.

Future research could address this issue in several ways. First, studies should include traditional measures of cognitive abilities so further data can be accumulated. Second, the constructs assessed by interviews should be examined at both conceptual and empirical levels. Evidence of the constructs underlying the criteria would also be helpful. Third, because of the availability and low cost of tests, interviews should be designed to complement rather than duplicate tests. For example, attributes such as interpersonal skills might be ideally measured in an interview where both verbal and nonverbal information can be judged (Ulrich & Trumbo, 1965). Finally, the effects of the level of interview structure on other components and the range of cognitive abilities in the sample should be considered when interpreting relationships with tests.

Another research and practice topic is whether some forms of job analyses are more useful for developing structured interviews. Many issues revolve around the popular critical incidents technique, such as how the information is turned into interview questions, whether unique insight can be gained using contrasting groups or the repertory grid, and whether behavioral consistency might lead to questions that are face valid but fakable. In practice, critical

incidents should be supplemented with information on job tasks and requirements (Feild & Gatewood, 1989; Langdale & Weitz, 1973; Wiener & Schneiderman, 1974). Finally, other methods of job analysis have not been extensively examined for the purpose of developing interview questions (e.g., protocol analyses, diaries, or questionnaires linked to interview items).

2. Ask Exact Same Questions of Each Candidate

Explanation and Alternatives. The most basic component of structure is standardization of questioning. It may be the first component that emerged in the literature, with early studies stipulating question content and sequence through such means as interview guides (Hovland & Wonderlic, 1939) and question patterns or arrays (McMurry, 1947). An early review defined a structured interview as "one conducted according to an established pattern" (Wagner, 1949, p. 29). Early authors suggested that the idea of structuring interviews for employment purposes was inspired by Binet's success using structured interviews for intelligence testing of children (Wagner, 1949; Wonderlic, 1942).

The range of alternatives on this component can be summarized by four levels of structure. These levels are similar to those used by Huffcutt and Arthur (1994), except here this component is separated from prompting and follow-up questioning (component 3). The first and highest level requires that the exact same questions be asked of each candidate in the exact same order (M. Campion et al., 1988, 1994; Delery et al., 1994; Edwards, Johnson, & Molidor, 1990; Green, Alter, & Carr, 1993; Hakel, 1971; Latham & Saari, 1984; Latham et al., 1980; Latham & Skarlicki, 1995; Reynolds, 1979; Robertson et al., 1990; M. K. Stohr-Gillmore, M. W. Stohr-Gillmore, & Kistler, 1990; Walters et al., 1993; Weekley & Gier, 1987).

The second highest level requires primarily that the same questions be asked, but allows some flexibility to tailor the interview to different candidates or to pursue interesting lines of discussion. These interviews may consist of lists of specific initial questions to use (Carlson, Thayer, Mayfield, & Peterson, 1971; Freeman, Manson, Katzoff, & Pathman, 1942; Hovland & Wonderlic, 1939; Mayfield, Brown, & Hamstra, 1980; McMurry, 1947), example

questions (Anderson, 1954), or arrays or patterns of questions to pick from (Janz, 1982; Nevo & Berman, 1994; Orpen, 1985) that are often organized by the construct assessed (Motowidlo et al., 1992; Pulakos & Schmitt, 1995).

The third level does not provide any actual questions. Instead, these interviews only provide outlines of topics to cover (Ghiselli, 1966; Yonge, 1956), lists of desirable candidate attributes or job requirements (Arvey, Miller, Gould, & Burch, 1987; Grove, 1981), or scales or forms to be filled out (Adams & Smeltzer, 1936; Barrett, Svetlik, & Prien, 1967; Carlson, Schwab, & Heneman, 1970). Other interviews in the literature are not described in sufficient detail to determine their structure precisely, but they also appear to provide only modest guidance (Campbell, Prien, & Brailey, 1960; Shaw, 1952).

The fourth and lowest level of structure is no structure. The interviewer is free to ask whatever questions he or she deems appropriate.

Hybrids are possible. For example, interviewers may pick questions from an array, but this would be done in advance and the same questions used for all candidates. An interview might combine required and discretionary questions or standard questions and a period of free questioning (Schuler, 1989).

Aside from whether the same questions are asked, it may be advantageous to sequence or group the questions within an interview. Organizing questions around rating dimensions, education or work history, or other logical systems may enhance the structure by simplifying the judgment process.

Effects on Reliability, Validity, and User Reactions. Interviews are essentially psychological tests and, as such, should be standardized samples of behavior. "Standardization implies uniformity of procedure in administration and scoring" (Anastasi, 1976, p. 25). Therefore, using the same questions may be the most basic requirement for converting the interview from a conversation into a scientific measurement.

This component may increase interrater and test-retest reliability of the content of the interview (Table 2) because different interviewers will not ask different questions, and the same interviewer will not ask different questions of different candidates. There may be less opportunity for interviewer-

candidate interactions with the same questions, and candidate consistency may increase because questioning will follow a controlled and predictable pattern.

Asking questions in the same order might further enhance the same types of reliability because it further increases the consistency of the interview. Also, grouping questions has been shown to increase internal consistency reliability (Schriesheim, Solomon, & Kopelman, 1989). This effect is not shown in Table 2 because it is not a result of using the same questions per se.

It was recognized early that standardized questioning might ease candidate comparisons and improve validity (Otis, 1944). The same questions will not ensure job-relatedness, but may reduce deficiency by not omitting questions (Table 2). It may reduce contamination by preventing discussion of tangential topics and other biasing influences on questions (Dipboye & Gaugler, 1993) and by reducing cognitive overload of the interviewer by focusing attention only on specific questions (Dipboye & Gaugler, 1993; Maurer & Fay, 1988). Grouping questions might further reduce deficiency and contamination by focusing the discussion on one topic at a time, as long as it does not enhance socially desirable responding by making the assessment dimensions obvious. Finally, meta-analytic reviews provide strong evidence for the reliability and validity benefits because they use this component as the primary defining characteristic of structured interviews (Conway et al., 1995; Huffcutt & Arthur, 1994; McDaniel et al., 1994; Wiesner & Cronshaw, 1988; Wright et al., 1989).

This component should reduce EEO bias and legal exposure because of the obvious fairness of asking all candidates the same questions (Table 2). It limits many of the sources of potential bias in the interview (Dipboye, 1994). Interviews with this component have been viewed as more defensible by attorneys (Latham & Finnegan, 1993), and it relates empirically to positive outcomes in court cases involving the interview (Gollub-Williamson et al., 1995).

This component may show both positive and negative effects on reactions, however. Candidates judge structured interviews as somewhat more face valid (Smither et al., 1993), but candidates may prefer the interactional freedom of unstructured interviews so they can display their credentials in the most

positive manner (Latham & Finnegan, 1993). A predetermined question order may also inconvenience candidates if they were prepared to describe their background in a different order (e.g., chronologically).

Likewise, interviewers may prefer the freedom of unstructured interviews in order to maintain power and control over decision making, communicate the organization's values, and qualitatively assess candidate fit (Dipboye, 1994). Conversely, interviewers may prefer a well-developed set of questions to enhance face validity, help them be organized and prepared for the interview, and objectively compare candidates (Latham & Finnegan, 1993). A question plan may give the interviewer a feeling of control and be a courtesy to candidates (Hakel, 1982). To date, research concerning reactions is ambiguous.

Research and Practice Issues. One question is what level of structure is needed on this component to attain high validity. Huffcutt and Arthur's (1994) meta-analysis indicates that structure beyond the second highest level has no additional value. But meta-analyses are limited by the number and variety of studies in the literature, and components of structure tend to be confounded. Their result contradicts the psychometric logic that asking exactly the same questions is more standardized than asking mostly the same questions.

If the same questions are not used, validity will be influenced by the questions selected. Future research might examine optimal strategies for selecting items during an interview. Potentially, concepts from Item Response Theory and adaptive testing could be used (Drasgow & Hulin, 1990). The idea would be to subsequently ask more difficult questions if the candidate correctly answered previous questions, and vice versa. The goal is to get an accurate measure of candidate aptitudes with the fewest number of questions.

Research is needed on user reactions. How much does standardized questions bother candidates? After all, it could enhance perceptions of fairness and interviewer preparation. Likewise, research is needed to determine whether interviewers see this component as helpful (Latham & Finnegan, 1993) or an unnecessary restriction of their freedom (Dipboye, 1994). Both parties may view the modest inconvenience as a worthwhile trade-off.

Past research on reactions has not been empirical (Dipboye, 1994) or methodologically strong. Often structured interviews are simply described to respondents and their reactions measured (Latham & Finnegan, 1993; Smither et al., 1993), rather than studying reactions in real selection contexts.

Finally, there are many practical issues. For example, questions that are used too often might be leaked to candidates. One solution is to have a large battery of questions from which to develop different interviews. The same set of questions could be used for a given hiring wave, but changed between waves. Equivalence could be established by using item analysis to equate question difficulty or, more simply, by selecting questions to equate content validity.

3. Limit Prompting, Follow-up Questioning, and Elaboration on Questions

Explanation and Alternatives. To many authors, the "essential character" of the interview is the "dynamic interaction between two people" (Yonge, 1956, p. 27). However, the use of prompts and follow-up questions is a primary means by which interviewers might bias information gathering (Dipboye, 1994).

The range of alternatives on this component can be summarized by four levels. The first and highest level is the prohibition of any prompting, follow-up questioning, or elaboration. If necessary, questions can be repeated, or candidates can be given a card containing the question (Green et al., 1993). This level tends to occur with those approaches that also use the exact same questions (M. Campion et al., 1988, 1994; Latham & Saari, 1984; Latham et al., 1980; Lowry, 1994; Walters et al., 1993; Weekley & Gier, 1987).

The second highest level is to only allow limited or pre-planned prompts and follow-up questions. For example, suggested probes might be provided that the interviewer can use if needed (Carlson, et al. 1971; Mayfield et al., 1980), only probes using the same wording as the assessment dimension may be used, interviewers can be allowed a specified number of follow-up questions (Freeman et al., 1942), or after each answer the interviewer can be allowed to simply ask, "Is there anything else you would like to add?" (Robertson et al., 1990).

The third level is to allow and even encourage the unlimited use of probes and follow-ups. Sometimes follow-ups and probes are required to get pertinent

information (McMurry, 1947), explore negative answers (Hovland & Wonderlic, 1939), keep the candidate on track and avoid evading questions (Janz, 1982; Orpen, 1985), test hypotheses about the candidate (Drake, 1982), seek disconfirming evidence before drawing conclusions (Green, 1995), or simply get a specific and detailed answer or otherwise satisfy the need for information (Green et al., 1993; Hakel, 1971; Motowidlo et al., 1992; Reynolds, 1979). In panel interviews, probing is often assigned to interviewers not asking the questions (Roth & J. Campion, 1992). Unlimited probing has been used as a distinguishing feature between semi-structured and structured interviews (Heneman, Schwab, Huett, & Ford, 1975; Schwab & Heneman, 1969).

The fourth and lowest level is no guidance to the interviewer on probing and follow-up. This is the most frequent level of structure on this component. This level might also include some clinically oriented interviews that attempt to go beyond the content of the answers to probe the underlying meaning of the candidate's responses (Kelly & Fiske, 1950).

Effects on Reliability, Validity, and User Reactions. Limiting prompting and follow-up questions is expected to have the same effects on reliability as using the same questions (Table 2). Because of variation between interviewers on the types and extent of prompts used, structuring may increase interrater reliability of the content. Because the same interviewer might use different prompts and follow-ups across candidates, test-retest reliability may increase, and interviewer-candidate interactions may decrease. Finally, candidate consistency might increase because questioning would be less spontaneous.

This component may have varied effects on validity (Table 2). Because prompting and follow-up questioning are intended to clarify answers and seek additional information, limiting their use might create deficiency (Huffcutt & Arthur, 1994). However, they could also create contamination if they introduce extraneous information into the interview, or if they coach candidates into giving the right answers. Therefore, the net effect on validity is unclear.

The effects on user reactions are also varied (Table 2). It may have a positive effect on reducing EEO bias because the potential for asking illegal

questions or showing favoritism through coaching is reduced. Conversely, this component may have negative effects on reactions. Candidates may prefer prompting and follow-ups because they enhance the conversational nature of the interview, thus allowing the flexibility to best highlight their credentials. Interviewers may also prefer no limitations so they can use their intuition to probe for useful information or influence the interview (Dipboye, 1994).

Research and Practice Issues. First, do prompting, follow-up questioning, and elaborating on questions lead to increased or decreased validity? Improved reliability and reduced contamination should increase validity, but validity could decrease if deficiency occurs. Second, how much of this component is needed? Follow-up questions that clarify confusion are not likely to have negative effects. However, prompting that influences candidate answers or changes the constructs assessed would be undesirable.

Rather than focusing on whether these limitations should be imposed, future research might focus on improving their use so that both deficiency and contamination are avoided. For example, standardized prompts and follow-ups linked to job requirements could reduce deficiency without creating contamination. Similarly, neutral prompts and follow-ups (e.g., "Please tell me more." "Could you explain further?" "Please provide an example.") would not change the construct assessed, yet could be standardized in their placement and frequency of usage (e.g., one after each question; Robertson et al., 1990). Prompts can also be coded and analyzed for their effects on interview content (Dillman, 1978), thus allowing their use and an assessment of biasing effects.

Finally, research is needed on the anticipated user reactions. Do these limitations actually make the interview too unnatural? Do they limit interviewer freedom unnecessarily? Would these limitations be accepted if their advantages in terms of consistency and fairness were explained?

4. Use Better Types of Questions

Explanation and Alternatives. Question type can refer to either how the question is asked or its content. For example, questions might ask for self-descriptions, reactions to hypothetical situations, or answers describing past

behaviors. Questions might ask about general background, specific knowledge or skills, motivation, or other constructs.

Unlike most other components, different types of questions cannot be neatly ordered in terms of increasing structure. However, certain types of questions are more structured than others for two reasons. First, some questions are more frequently used with high levels of structure on other components (e.g., situational questions tend to use the same questions and limit prompting). Second, some types of questions are more specific because of their focused nature (e.g., asking for past behaviors to support answers). Therefore, question structure is only distinguished in a general way, based largely on the types that have been empirically examined.

One type that has been widely studied and is considered relatively structured is situational questions (M. Campion et al., 1988, 1994; Delery et al., 1994; Freeman et al., 1942; Hakel, 1971; Latham & Saari, 1984; Latham et al., 1980; Latham & Skarlicki, 1995; Robertson et al., 1990; Schmitt & Ostroff, 1986; Stohr-Gillmore et al., 1990; Walters et al., 1993; Weekley & Gier, 1987). They pose hypothetical situations that may occur on the job, and the candidates are asked what they would do. Situational questions are usually used with high levels of structure on other components, including job analysis, same questions, limited prompting, and ones yet to be discussed. Also, situational questions tend to be highly specific, which further enhances their structure.

A second widely studied and fairly structured type is past behavior questions (Green et al., 1993; Grove, 1981; Janz, 1982; Motowidlo et al., 1992; Orpen, 1985; Pulakos & Schmitt, 1995). In contrast to situational questions that focus on future behavior, these focus on past behavior by asking candidates to describe what they did in past jobs as it relates to requirements of the job they are seeking. A slight variant asks for past accomplishments to support the answers (Tarico, Altmaier, Smith, Franken, & Berbaum, 1986) based on research using accomplishment records for personnel selection (Hough, 1984). Past behavior questions are usually used with moderately high structure on other components (e.g., job analysis, similar questions across interviews), and

their highly specific nature further enhances structure.

A third fairly structured type is background questions. They focus on work experience, education, and other qualifications (Carlson et al., 1971; Lopez, 1966; Mayfield et al., 1980; Roth & J. Campion, 1992). Early interviews often asked about family and personal background (Hovland & Wonderlic, 1939; McMurry, 1947), which is illegal or ill-advised today.

A fourth relatively structured type is job knowledge questions. Interviews in the literature have tended to mix these questions with other types (Arvey et al., 1987; M. Campion et al., 1988; Walters et al., 1993). In addition to asking for documentation of job knowledge, these questions may ask candidates to actually demonstrate specific job knowledge. The highly specific nature of these questions enhances structure.

These four types of questions are illustrated in Table 3. Parallel examples are given to allow for comparison.

Other question types also lend themselves to structured interviews, such as job samples or simulations that present actual job tasks, or "willingness" questions that query candidate understanding of aversive job requirements (e.g., travel, shift work). No research has focused uniquely on these types, but they have been included in structured interviews (Campion et al., 1988).

Many question types could be considered relatively less structured because they tend to be vague. Examples include questions on opinions and attitudes, goals and aspirations, and self-descriptions and self-evaluations. They are sufficiently ambiguous to allow candidates to present their credentials in an overly favorable manner or to subvert questions they cannot answer. They also tend to focus on poorly defined traits with uncertain links to job performance.

Effects on Reliability, Validity, and User Reactions. Comparative studies have confounded question type with other components of structure, such as job analysis and same questions (Heneman et al., 1975; Janz, 1982; Maurer & Fay, 1988; Orpen, 1985; Schwab & Heneman, 1969), thus the effect of question type on reliability has not been examined in isolation. Better questions may increase reliability if their specific and unambiguous nature enhances consistency in

the candidates, or if better and more similar questions have higher internal consistency (Table 2). However, there is little evidence at this time on the effects of question type on reliability, so these expectations are speculative.

The most likely effect is enhanced validity (Table 2). This should occur through enhanced job-relatedness and by reducing contamination from low quality questions. There are also theoretical arguments and empirical evidence. Situational questions may predict because of the well-established relationship between goals or intentions and future behavior (Locke & Latham, 1984). Past behavior questions may predict because of the well-established axiom that past behavior predicts future behavior. Also, past behavior questions in the form of biodata instruments have shown good validity (Mumford & Stokes, 1992). Most of the articles on situational and past behavior questions reviewed above present validity evidence. Job knowledge questions are supported by validity generalization evidence (J. Hunter & R. Hunter, 1984). Job sample and simulation questions may predict because they are samples rather than signs of behavior (Wernimont & Campbell, 1968), and willingness questions may predict based on evidence supporting realistic job previews (Wanous, 1980).

Better questions may positively influence user reactions (Table 2). Reduced EEO bias is expected due to the preliminary test fairness evidence for both situational (M. Campion et al., 1988) and past behavior questions (Pulakos & Schmitt, 1995). Interviewer reactions may be positive if they believe that the questions enable them to make better decisions (Latham & Finnegan, 1993).

Research and Practice Issues. The search for better question types has been a popular topic in recent years. Some studies have compared questions. For example, M. Campion et al. (1994) compared future (situational) versus past (behavior) questions in a highly structured interview in an actual selection context in a pulp mill. Both types had empirical validity against supervisor ratings, but past questions were more valid. Similarly, Pulakos and Schmitt (1995) compared situational versus past behavior questions in a fairly structured interview in a research study of law enforcement employees. They found that only past behavior questions were valid against supervisor ratings.

Conversely, Latham and Saari (1984) compared situational versus past behavior questions in a research study of clerical personnel and found that only situational questions were valid against supervisory and peer rating criteria. Maurer and Fay (1988) similarly compared situational versus past behavior questions in a laboratory study and found that situational questions had higher interrater agreement. Note that both studies operationalized past behavior questions in terms of fairly broad inquiries about past experiences and training, rather than specific questions that required candidates to give specific examples of past experiences. Latham and Skarlicki (1995) operationalized past behavior questions correctly in a comparison with situational questions in a research study of college faculty, and they again found only situational questions were valid against peer rating criteria.

Furthermore, McDaniel et al. (1994) compared 16 validity coefficients for situational with 127 coefficients for job-related interviews and found situational interviews slightly more valid. However, this does not address the issue well because the "job-related" interviews were very heterogeneous, including many other question types in addition to past behavior. There were too few coefficients to analyze past behavior interviews alone.

In summary, validity superiority between situational and past behavior questions cannot be determined from the current evidence. Differences in validity may be due more to the constructs assessed than to question type. The constructs that can be appropriately assessed with different question types should be examined in future research.

Aside from validity, future research might compare other dimensions. For example, situational questions might work better with candidates lacking sufficient work experience to respond to past behavior questions. Previous comparative studies have tended to use only experienced candidates. Also, past behavior questions may be less fokable than situational questions due to their potentially verifiable nature, but this has not been tested.

There are also subtle nuances in using these questions that may be very important. For example, past behavior questions may require highly specific

responses to be valid (Green et al., 1993), and situational questions may require posing a dilemma to reduce the potential for faking (Latham & Skarlicki, 1995) as illustrated in the first example in Table 3.

The influence of question type on reliability is unresolved. Unlike past studies, future comparisons of reliability should be certain to hold other components of structure constant. Finally, future research should explore other question types, such as simulations and willingness questions.

In practice, it is not likely that the sole use of just one question type is desirable. As long as different types have adequate validity, a range of questions offers variety for both the candidate and interviewer.

5. Use Longer Interview or Larger Number of Questions

Explanation and Alternatives. Length is a basic, but overlooked, component of structure. Within reasonable limits, longer interviews are more structured because they obtain a larger amount of information. Length can be reflected in either administration time or number of questions.

Surprisingly, many articles do not report this essential information. Of those that do, a very wide range is seen. The 38 studies reporting time range from 3 to 120 minutes, with a mean of 38.95 ($SD = 25.79$). The 14 reporting the number of questions range from 4 to 34, with a mean of 16.50 ($SD = 8.71$). No articles explain the reasoning behind the choice of interview length.

Effects on Reliability, Validity, and User Reactions. The most direct effect of length is on internal consistency reliability (Cronbach, 1951). Length may indirectly affect test-retest and interrater reliability of the evaluation process because longer interviews may produce more stable and comparable measurements (Table 2). Length should affect validity because longer measures should be less deficient. Longer interviews may be more important for higher quality candidates (Tullar, Mullins, & Caldwell, 1979), perhaps because there is more information to evaluate.

Paradoxically, Marchese and Muchinsky (1993) found interview length negatively related to validity ($r = -.29$). However, this may be spurious because it was only marginally significant ($p < .10$), was not well explained,

and did not control for the other components of structure. Yet, overly long interviews may collect too much information such that overload occurs, and decision quality is reduced (Dipboye, Fontenelle, & Garner, 1984; Oskamp, 1965).

Finally, longer interviews could elicit somewhat negative reactions from candidates and interviewers because they are more effortful to complete.

Research and Practice Issues. Psychometrics are probably adequate to understand the theoretical influences of interview length, thus the issues are related to practice. What are the reasonable upper and lower limits? Probably interviews that exceed an hour tax the patience of participants, and very brief interviews may be inadequately reliable. Two-thirds of the interviews in the literature are between 30 and 60 minutes, and half contain 15 to 20 questions.

6. Control Ancillary Information

Explanation and Alternatives. A considerable threat to structure is the uncontrolled use of ancillary information. This includes application forms, work histories, test scores, recommendations, results of previous interviews, transcripts, personnel files, and so on. Many interviews have considered ancillary information (Albrecht, Glaser, & Marks, 1964; Bobbitt & Newman, 1944; Bolanovich, 1944; Dougherty, Ebert, & Callender, 1986; Grove, 1981; Hakel, 1971; Handyside & Duncan, 1954; Harris, 1972; Holt, 1958; Huse, 1962; Kelly & Fiske, 1950; Putney, 1947; Raines & Rohrer, 1955; Roth & J. Campion, 1992; Shaw, 1952; Trankell, 1959; Waldron, 1974).

With exceptions (Roth & J. Campion, 1992), these interviews had several things in common. First, they were among the more unstructured interviews. Second, they were often conducted by psychologists as part of individual assessments. Third, sometimes the studies were embedded in the debate between clinical versus statistical prediction during the 1950s and 1960s (Meehl, 1954, Sawyer, 1966). Fourth, sometimes the interviews were in an assessment center where many types of information were collected. Last, often the article was unclear as to precisely what information was available to interviewers.

Two problems are created by ancillary information. First, it confounds the interpretation of the value of the interview. It is uncertain if validity

is due to information gained in the interview or to this other information (Ulrich & Trumbo, 1965; Webster, 1959). Second, it creates unreliability if the same information is not available for all candidates or not given to all interviewers, or if interviewers evaluate the information differently.

To enhance structure, ancillary information can be withheld from the interviewers (M. Campion et al., 1988; Latham et al., 1980). This is not to suggest that such information should be excluded from consideration. Instead, it should be used as a separate predictor, or the interview should be structured so the information is available for all candidates and evaluated in a standardized manner (Carlson et al., 1971; Mayfield et al., 1980).

Effects on Reliability, Validity, and User Reactions. Either withholding this information or standardizing it should increase the test-retest and interrater reliability of the interview (Table 2). There is even some evidence to this effect (Dipboye et al., 1984).

The effects on validity are uncertain, however. Withholding this information could reduce contamination if it is not valid or is used incorrectly, but withholding it could increase deficiency if it is valid. In supplementary analyses of 66 validity coefficients, McDaniel et al. (1994) found that average validities were higher when test information was not available, and this result held even when interview structure was controlled.

This result seems counter to the expectation that such information could make the interview more complete (Tucker & Rowe, 1977) or the validity support for some types of information such as tests (J. Hunter & R. Hunter, 1984). Conversely, test information can be used unreliably, prescreening on ancillary information can cause restriction of range, and the effect may depend on the information (Dalessio & Silverhart, 1994). Also, most studies are ambiguous regarding the availability of ancillary information, thus meta-analyses must be viewed with caution. Yet, the weight of evidence suggests that standardizing, if not withholding, ancillary information is advised.

The effects on user reactions may be mixed. EEO bias might be reduced if it prevents the consideration of personal, family, social, or other information

unrelated to the job. However, interviewers may react negatively to not having access to relevant information before the interview. Likewise, candidates may react negatively to an interviewer who is unaware of relevant information that was submitted on applications and resumes before the interview.

Research and Practice Issues. Future research should determine if this component has the expected effects on reliability and validity and how such information can be standardized.

A practical issue is how to avoid potential negative reactions. It may be sufficient to simply explain. Another practical issue is how to make sure critical background information is complete (e.g., education and employment, address and phone, etc.) if it is not verified in the interview. One solution is to have a clerk verify the information.

7. Do Not Allow Questions from Candidate Until After the Interview

Explanation and Alternatives. Candidates naturally have many questions about the job and organization (e.g., pay and benefits, working conditions, etc.). Selling the organization to the candidate is part of some interviews. However, uncontrolled questions from the candidate reduces standardization by changing the interview content in unpredictable ways. Thus, structure can be enhanced by not allowing candidate questions during the interview. Instead, time should be allowed outside the interview for these questions.

Most articles do not mention this component. The notion in unstructured interviews is that the atmosphere is conversational, with both parties asking questions (Drucker, 1957). Some articles even note that the interview provides valuable information for the candidate (Komives, Weiss, & Rosa, 1984). Only one article mentioned this component by stating that candidates were allowed to ask questions in a later, nonevaluation interview (M. Campion et al., 1988).

Effects on Reliability, Validity, and User Reactions. Not allowing questions from candidates should help standardize the content, thus increasing test-retest and interrater reliability (Table 2). Reducing the conversational nature should enhance candidate consistency and reduce effects of interviewer-candidate interactions. However, the effects on validity may be mixed. It

could decrease contamination, but it could also increase deficiency if it precludes relevant information from emerging. Finally, not allowing candidate questions could elicit negative candidate and interviewer reactions because it restricts their freedom and may lead to an awkward conversation. It also prevents interviewers from using the nature of candidate questions to judge candidates, and it precludes candidates from taking the popular advice to ask questions and use the information to shape their answers (Beatty, 1986).

Research and Practice Issues. First, is the net effect on validity positive or negative? Is the potential increase in reliability and decrease in contamination more important than the potential increase in deficiency? Second, are candidate and interviewer reactions sufficiently negative such that this component of structure should not be used?

A compromise is to allow questions that clarify ambiguity, but not questions that are tangential. Further, ample time can be allowed afterward to answer questions, and this should be made known in advance. If the informal chattiness of the interview is reduced, it may be justified because it is a focused conversation with a specific purpose. Another possibility would be to provide a separate meeting for answering questions. For example, organizations often conduct socials with college candidates to provide company information.

8. Rate Each Answer or Use Multiple Scales

Explanation and Alternatives. There are two elements to this component. First, ratings can be made as the candidate is responding or at the end. Second, multiple ratings or only a single global rating can be made. Rating during the interview is more structured because judgments are more directly linked to specific responses. Multiple ratings are more structured because they are more detailed and thorough. These elements are considered together because they define three commonly observed levels of structure. These levels are similar to those suggested by Huffcutt and Arthur (1994).

The first and highest level is to rate each answer as it is given, often with scales tailored to each question. This is used by the most structured approaches (M. Campion et al., 1988; Delery et al., 1994; Latham & Saari, 1984;

Latham et al., 1980; Latham & Skarlicki, 1995; Weekley & Gier, 1987) and some fairly structured approaches (Green et al., 1993; Motowidlo et al., 1992).

The second level is to make multiple ratings at the end. Ratings are made on dimensions or scales, ranging from 2 to 12 or more, based on answers to multiple questions or on the entire interview. This level is less structured than rating every answer because judgments are not as linked to individual answers, and there are fewer scales than questions so fewer ratings are made.

Due to flexibility, this level is used in interviews that span the range of structure on other components, including highly structured (Robertson et al., 1990; Walters et al., 1993), moderately structured (Arvey et al., 1987; Barrett et al., 1967; Borman, 1982; Grove, 1981; Hakel, 1971; Janz, 1982; Landy, 1976; Mayfield et al., 1980; Orpen, 1985; Pulakos & Schmitt, 1995; Reynolds, 1979; Roth & J. Campion, 1992; Yonge, 1956; Zedeck et al., 1983), and fairly unstructured (Anderson, 1954; Bolanovich, 1944; Campbell, 1962; Dougherty et al., 1986; Drucker, 1957; DuBois & Watson, 1950; Fisher et al., 1967; Freeman et al., 1942; Glaser, Schwarz, & Flanagan, 1958; Hilton et al., 1955; Hovland & Wonderlic, 1939; Huse, 1962; Komives et al., 1984; Maas, 1965; Morse & Hawthorne, 1946; Rafferty & Deemer, 1950; Raines & Rohrer, 1955; Reeb, 1969; Shaw, 1952; Trankell, 1959; Tubiana & Ben-Shakhar, 1982; Waldron, 1974).

The third level is to make one overall judgment or prediction at the end of the interview. This level is typical, but not unique, to less structured and older approaches (Campbell et al., 1960; Ghiselli, 1966; Harris, 1972; McMurry, 1947; Meyer, 1956; Mischel, 1965; Plag, 1961; Pulos, Nichols, Lewinsohn, & Koldjeski, 1962). Many interviews that make ratings on multiple dimensions will also make an overall rating at the end. Similarly, sometimes the candidates will be ranked (Carlson et al., 1970; Schwab & Heneman, 1969).

Effects on Reliability, Validity, and User Reactions. Rating each answer should increase test-retest and interrater reliability because each rating is based on the response to the same question (Table 2). With ratings conducted at the end, ratings of different candidates (or by different interviewers) on a given scale may be based on different questions. Also, reliability may be

increased because rating each answer as it is given is less cognitively complex than rating a set of dimensions at the end, both in terms of memory requirements (because ratings are done immediately) and processing requirements (because ratings are based on single behavioral events). There is also some evidence that decomposed judgments are more reliable than holistic judgments (Armstrong, Denniston, & Gordon, 1975; Butler & Harvey, 1988; Einhorn, 1972; but cf. Cornelius & Lyness, 1980). Finally, moving up levels of this component typically means making more ratings, thus internal consistency may increase. The meta-analysis by Conway et al. (1995) supports the strong reliability benefits of making multiple ratings.

This component should increase validity. Deficiency is reduced because more behaviors are evaluated. With specific scales, contamination is reduced because only relevant behaviors are evaluated. Also, the construct validity of rating individual questions may be better than rating dimensions across questions because of the assessment center findings that ratings reflect exercises more than dimensions (Sackett & Dreher, 1982).

This component should not influence user reactions unless ratings are made in an obvious manner that creates evaluation apprehension for the candidates, or unless interviewers feel rushed making ratings during the interview.

Research and Practice Issues. Minor practical difficulties rating each question can be overcome. For example, dimension subscores are often desired. They can help match candidate attributes to job requirements, provide feedback to candidates, or understand results at a detailed level. One solution is to rate each question, then sum the questions that bear on a given dimension. If dimension ratings are desired, questions on the same dimension can be clustered together or clearly indicated (Schmitt & Ostroff, 1986) to reduce complexity. Also, rather than wait until the end, ratings can be made during the interview as questions on each dimension are completed.

In addition, developing a customized rating scale for each question can be cumbersome, but this can be handled by using a common rating scale for all questions. Finally, different questions are sometimes asked across interviews,

but each question can still be rated and a simple average calculated.

9. Use Detailed Anchored Rating Scales

Explanation and Alternatives. Like standardizing questions, this was an early component of structure to emerge (Adams & Smeltzer, 1936; Freeman et al., 1942; Smeltzer & Adams, 1936). Enhancements to interview rating scales reflect developments in measurement in other areas. For example, after extensive research on anchored scales for performance appraisal in the 1970's (Landy & Farr, 1980), subsequent published interviews nearly all have anchored scales.

Anchored rating scales use behavioral examples to illustrate the levels to make them less ambiguous and to eliminate semantic differences possible with adjective anchors (Smith & Kendall, 1963). Modeled after Thurstone (1927), developing such scales involves collecting actual candidate answers, judging the goodness of the answers by experts, and selecting unambiguous answers to illustrate points along the scale. Simpler procedures involve intuitively developing examples for good, average, and poor answers on a scale.

At least four types of anchors have been used (Table 4). First, anchors can be example answers or illustrations. They might not be the exact words candidates use but only examples they might be expected to use (Smith & Kendall, 1963). Second, anchors can be descriptions or definitions of answers. Here, the quality of the answer is described narratively, rather than in terms of potential candidate words. These anchors avoid the tendency of interviewers to look for exact matches with the example answers. Third, anchors can contain evaluations of the answers (e.g., excellent, good, poor). Fourth, anchors can contain relative comparisons (e.g., answer given by the top 20% of candidates).

There are four levels of structure. The highest level uses multiple types of anchors (Anderson, 1954; M. Campion et al., 1988, 1994; Green et al., 1993).

The second highest level uses primarily a single type of anchor. They usually use either example answers (Edwards et al., 1990; Hakel, 1971; Latham & Saari, 1984; Latham et al., 1980; Latham & Skarlicki, 1995; Lowry, 1994; Maas, 1965; Nevo & Berman, 1994; Schmitt & Ostroff, 1986; Vance, Kuhnert, & Farr, 1978; Weekley & Gier, 1987) or descriptions of answers (Campbell, 1962; Grove,

1981; Janz, 1982; Motowidlo et al., 1992; Pulakos & Schmitt, 1995; Robertson et al., 1990; Stohr-Gillmore et al., 1990; Tarico et al., 1986; Zedeck et al., 1983) although the anchors may include evaluative words as well.

The third level uses either unanchored scales or numbers or adjectives as anchors. On average, these are older interviews (Arvey et al., 1987; Barrett et al., 1967; Bender & Loveless, 1958; Dougherty et al., 1986; DuBois & Watson, 1950; Fisher et al., 1967; Holt, 1958; Huse, 1962; Komives et al., 1984; Landy, 1976; McMurry, 1947; Mischel, 1965; Morse & Hawthorne, 1946; Reeb, 1969; Reynolds, 1979; Shahani, Dipboye, & Gehrlein, 1991; Walters et al., 1993).

The fourth level does not require quantitative judgments. Instead, they use written summaries (Bobbitt & Newman, 1944), relative rankings (Albrecht et al., 1964; Carlson et al., 1971), or group discussion (Flynn & Peterson, 1972; Gardner & Williams, 1973; Handyside & Duncan, 1954; Kelly & Fiske, 1950).

There are other approaches that do not fit into the above typology. For example, graphic scales can be enhanced by rating a large number of behavioral statements (Campbell et al., 1960; Drucker, 1957; Mayfield et al., 1980). Any scale can be enhanced by providing a detailed description of the underlying dimension (Campbell, 1962; Dougherty et al., 1986). Other scaling approaches used with interviews include checklists (Hovland & Wonderlic, 1939; Raines & Rohrer, 1955), forced choice items (Drucker, 1957), forced distributions (Fisher et al., 1967), and stanine scales (Trankell, 1959).

Effects on Reliability, Validity, and User Reactions. Anchored rating scales are presumed to enhance objectivity; thus, they are expected to increase test-retest and interrater reliability and interrater agreement (Table 2). If enhanced objectivity increases accuracy, they may also reduce contamination and deficiency. Because anchors are usually in terms of job behaviors or attributes, they also might increase job relatedness.

Two studies found that anchored scales have higher interrater reliability than unanchored scales in an interview (Maas, 1965; Vance et al., 1978), and one study also found higher accuracy (Vance et al., 1978). However, more extensive research on performance appraisal does not unambiguously support the

value of anchored rating scales over simpler scales (Landy & Farr, 1980).

Anchored scales should positively influence user reactions. The enhanced objectivity and preplanned nature of anchoring should reduce EEO bias (Arvey & Faley, 1988). Gollub-Williamson et al. (1995) found that behavioral criteria in interviews were related to positive court decisions. Anchored scales may also enhance interviewer reactions by easing the difficulty of judging answers.

Research and Practice Issues. The logical appeal of such scales may be contributing to their popularity, rather than strong evidence. Thus, future research must clearly determine their effects on reliability and validity.

Just as past studies borrowed from performance appraisal research in the 1970s, future studies might borrow from appraisal research in the 1980s. For example, that research suggests supervisors remember their past conclusions about employee performance better than the facts about performance (Murphy & Cleveland, 1995). Does this suggest ratings should be made during the interview before specific answers are forgotten? The research also discovered that cognitive schemas or mental models may influence judgments. Should interviews be structured to elicit, change, or capitalize on these schemas?

A pragmatic issue is how to develop anchors. Several methods have been suggested. Example answers could be obtained from actual candidates using pilot interviews for this purpose. Records of answers to similar questions from previous candidates could be used (Green et al., 1993). Interviewers could be asked about answers given by previous candidates (Latham et al., 1980). Brainstorming could be conducted with job experts (M. Campion et al., 1988), such as incumbents and supervisors (Robertson et al., 1990; Weekley & Gier, 1987) and personnel representatives.

10. Take Detailed Notes

Explanation and Alternatives. Notetaking is presumed to enhance interview structure because it reduces memory decay (M. Campion et al., 1988) and helps avoid recency and primacy effects (Schmitt & Ostroff, 1986). These effects may be most apparent when ratings are made at the end or when ratings are based on multiple questions. Notetaking requires justifying or documenting the rating

given. This encourages interviewers to attend to candidate answers in more detail and to organize their thoughts, thus possibly increasing accuracy.

There are several distinctions in notetaking with implications for the degree of structure. First, notetaking can be extensive or brief. Obviously, more extensive notetaking, such as summarizing each answer, is more structured. Second, notetaking can be required or optional, with required being more structured. Third, notetaking can record candidate answers or facts about credentials, or it can record evaluations or judgments. Recording answers and facts is more structured because it helps ensure that information is perceived accurately before judgments are made. Fourth, notetaking can occur during or after the interview, with during more structured due to less memory loss.

Three levels of structure can be seen in the literature. The highest level involves extensive, required notetaking of answers during the interview (M. Campion et al., 1988, 1994; Green et al., 1993; Janz, 1982; Latham et al., 1980; Mayfield et al., 1980; Motowidlo et al., 1992; Orpen, 1985; Pulakos & Schmitt, 1995; Robertson et al., 1990; Roth & J. Campion, 1992). The next level involves optional or brief notes of either answers or evaluations, often only at the end of the interview (Drucker, 1957; Heneman et al., 1975; Shaw, 1952; Tarico et al., 1986; Yonge, 1956). The lowest level is no notetaking.

Also related to structure, some studies developed special forms to record notes. For example, Grove (1981) had an "evidence organizer," Handyside and Duncan (1954) used record sheets, and McMurry (1947) used a biographical data form. The independent effects of such forms are impossible to assess because they are confounded with other components, like using the same questions.

Notetaking probably occurred in other interviews, but was not considered important enough to the researcher to mention. However, notetaking is very important to interviewers. In fact, it has been observed that interviewers may concentrate more on providing detailed notes than on making accurate ratings, much to the chagrin of the researcher (Meyer, 1956).

Effects on Reliability, Validity, and User Reactions. If interviewers take detailed notes, evaluation should be more consistent, thus increasing

test-retest and interrater reliability (Table 2). There should be less disagreement on what candidates said, thus higher interrater agreement.

There is evidence that notetaking can enhance recall in interviews (Macan & Dipboye, 1994; Schuh, 1980). It is analogous to using diaries in performance appraisal, and research suggests that diaries help organize information and increase accuracy (DeNisi, Cafferty, & Meglino, 1984; DeNisi, Robbins, & Cafferty, 1989). There is extensive research in education showing that the process of notetaking may increase learning, and being able to review notes later is also very important to learning (Carrier & Titus, 1979; Kiewra et al., 1991). There is some research showing notetaking helps jurors make correct distinctions (ForsterLee, Horowitz, & Bourgeois, 1994). All this suggests notetaking should increase accuracy and, perhaps, job-relatedness. Furthermore, notetaking should reduce deficiency because it helps ensure that important information is recorded and considered, and it should reduce contamination because of its verifiable nature.

Notetaking could have a mixed influence on user reactions. It may reduce EEO bias by focusing interviewer attention on candidate answers and away from illegal factors. Being able to reconstruct answers provides documentation needed for legal defensibility (Pursell et al., 1980), and keeping records of interviews is related to positive court decisions (Gollub-Williamson et al., 1995). However, interviewers may not be positive if requirements for detailed notes are burdensome. Notetaking has been found to be distracting in a clinical interview (L. Hickling, E. Hickling, Sison, & Radetsky, 1984). Candidate reactions are ambiguous. Notetaking may reduce eye contact, increase evaluation apprehension, and decrease conversational naturalness (Dipboye & Gaugler, 1993). Some evidence suggests people prefer counselors who refrain from notetaking (Miller, 1992). Nevertheless, it shows that the interviewer is paying attention to the answers and the answers are important enough to record.

Research and Practice Issues. There is inadequate evidence on the effects of notetaking on reliability and validity. Research should also examine the processes through which notetaking works, such as organizing information,

reducing memory decay, or abating rating errors. It is also ambiguous whether notetaking hurts candidate reactions and how such drawbacks can be avoided.

Video or audio taping could be used in place of notetaking, but reviewing tapes is time consuming, and taping does not elicit the cognitive processing notetaking requires. Taping could provide a record for defensibility, however, and its monitoring function may motivate the interviewer to administer the interview conscientiously. Having raters evaluate tapes may enhance structure by minimizing effects of interactions between candidates and interviewers. This might be cost effective for screening interviews if it reduces travel.

11. Use Multiple Interviewers

Explanation and Alternatives. Prior reviews have described the use of multiple interviewers, such as the panel interview, as a "promising" approach for improving reliability and validity (Mayfield, 1964, p. 252; Arvey & J. Campion, 1982, p. 293). Multiple interviewers may be beneficial for several reasons. Sharing different perceptions may help interviewers become aware of irrelevant inferences they make about variables that are not job-related (Arvey & J. Campion, 1982). Multiple interviewers may reduce the impact of idiosyncratic biases among individual interviewers (M. Campion et al., 1988; Hakel, 1982), and aggregating multiple judgments balances out random errors (Dipboye, 1992; Hakel, 1982). Recall of information may be better with multiple interviewers (Stasser & Titus, 1987). The range of information and judgments that different interviewers contribute may lead to more accurate predictions (Dipboye, 1992). Finally, more interviewers is akin to a longer test, thus the combined score should be more reliable (Hakel, 1982).

There are two distinctions here. First, multiple interviewers can conduct interviews together, or they can conduct interviews separately. The former is called a panel, board, or group interview, while the latter could be called a "serial" interview (Dipboye, 1992, p. 211). The former is more structured because all the interviewers hear the same candidate responses, while the same questions are typically not asked in serial interviews to avoid repetition. Second, the number of interviewers influences structure, with larger numbers

considered more structured. The upper range in the literature is five (DuBois & Watson, 1950) and nine (Hakel, 1971), with two or three most common.

Based on the first distinction, three levels of structure appear in the literature. The highest level is the panel interview (M. Campion et al., 1988, 1994; Drucker, 1957; DuBois & Watson, 1950; Edwards et al., 1990; Flynn & Peterson, 1972; Freeman et al., 1942; Glaser et al., 1958; Green et al., 1993; Landy, 1976; Latham & Saari, 1984; Latham et al., 1980; Latham & Skarlicki, 1995; Lowry, 1994; Morse & Hawthorne, 1946; Nevo & Berman, 1994; Pulakos & Schmitt, 1995; Reynolds, 1979; Roth & J. Campion, 1992; Stohr-Gillmore et al., 1990; Vernon, 1950). The next level is multiple interviews conducted separately (Bobbitt & Newman, 1944; Borman, 1982; Dougherty et al., 1986; Gardner & Williams, 1973; Handyside & Duncan, 1954; Hilton et al., 1955; Huse, 1962; Trankell, 1959). The lowest level is one interviewer.

It is noteworthy that this component is not strongly associated with other components of structure. Interviews that are unstructured on most other components may still use multiple interviewers. Second, a disproportionately large number of the studies on panels are in public sector settings. Selecting police is the most common (DuBois & Watson, 1950; Flynn & Peterson, 1972; Freeman et al., 1942; Landy, 1976; Lowry, 1994; Reynolds, 1979), but they have been used for a wide range of other civil service jobs (Bobbitt & Newman, 1944; Glaser et al., 1958; Morse & Hawthorne, 1946; Pulakos & Schmitt, 1995; Stohr-Gillmore et al., 1990; Vernon, 1950) including the military (Borman, 1982; Drucker, 1957; Gardner & Williams, 1973). This may be due to a heightened need for fairness perceptions when staffing government jobs.

Effects on Reliability, Validity, and User Reactions. Because multiple interviewers judge each candidate, total scores are based on multiple raters. Such scores should be more reliable than those based on single raters (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). If interviewers are exposed to the exact same answers (e.g., as in a panel), interrater agreement should be higher as well. Internal consistency should be higher because a greater number of judgments make up the total scores. Also, the potential for interviewer-

candidate interactions should be diluted with multiple interviewers.

Deficiency should be reduced because relevant information is less likely to be missed or overlooked with multiple interviewers. Contamination should be reduced because interviewers provide a check on each other to ensure irrelevant information does not enter the decision (Arvey & J. Campion, 1982).

Meta-analyses have examined this component. Conway et al. (1995) found that panel (versus separate) interviews were correlated .56 with reliability. Wiesner and Cronshaw (1988) found that among unstructured interviews, panels were more valid than individual interviews (.21 versus .11, or .37 versus .20 corrected). But among structured interviews, there was no real difference (.33 versus .35, .60 versus .63 corrected). McDaniel et al. (1994) found no difference between panel and individual formats among unstructured interviews (.18 versus .18, .33 versus .34), but a slight advantage for individual formats among structured interviews (.20 versus .25, .38 versus .46 corrected). Finally, Marchese and Muchinsky (1993) found no correlation between number of interviewers and validity.

Therefore, equivocal evidence exists for the validity benefits of multiple interviewers. However, there are three caveats. First, many panels were unstructured on other components, thus reducing their validity. The meta-analyses controlled for these other components in only a very gross way (e.g., structured versus unstructured dichotomy). Second, the preponderance of public sector settings for panel interviews may have an unknown confounding effect on validities. Third, there can be process losses with groups (e.g., conformity, conflict, domination, social loafing, etc.; Steiner, 1972), which may reduce the advantage of using multiple interviewers in some applications.

Multiple interviewers may reduce EEO bias if they reduce the effects of idiosyncratic biases. Having a system where decisions are reviewed by other interviewers has been related to positive court outcomes (Gollub-Williamson et al., 1995). Also, multiple interviewers allow members of different races or sexes to be represented, thus enhancing perceptions of fairness (Hakel, 1982). It is a common practice to ensure diversity on interviewing teams.

On the other hand, panels can be stressful for candidates. Stress may occur if panel members ask questions too quickly and overload the candidate, and the mere presence of multiple interviewers may enhance evaluation apprehension. This may occur intentionally when stress tolerance is a job requirement, such as in police work (Freeman et al., 1942). Either way, candidate reactions might be negative if panels are stressful.

Research and Practice Issues. One central question is whether multiple interviewers have an appreciable positive effect on reliability and validity. If yes, is it due to a longer measure, or do different interviewers evaluate different aspects of candidates? Also, are multiple interviewers needed when structure is high on other components? They may be most helpful when structure is moderate or judgments are difficult. Finally, research is needed on candidate reactions to this component.

There are many practical issues. For example, what is the best role for panel members? Should questions be asked by the same member, or should they rotate? Should all members take notes, or should one member run the interview and have the others take notes? Who should ask follow-up questions? Might panels be useful with self-managed work teams? They could allow team involvement yet ensure that the selection system is valid.

12. Use Same Interviewer(s) Across All Candidates

Explanation and Alternatives. Using the same interviewer is very important when other components are unstructured because different interviewers ask different questions, evaluate answers differently, consider different ancillary information, and so on. With different interviewers, there is no way to distinguish variance due to rating tendencies among interviewers (e.g., leniency/severity) from true score variance among candidates.

Dreher, Ash, and Hancock (1988) provided a series of arguments for how aggregating across interviewers might underestimate validity. First, nearly all studies combine interviewer data, thus concealing the problem. Second, interviewers make constant rating errors (e.g., leniency/severity). Third, interviewers differ in validity. Finally, validities aggregated across

interviewers are lower than individual validities due to these constant errors.

Several studies have found differences in validities between interviewers (Dipboye, Gaugler, & Hayes, 1990; Gehrlein, Dipboye, & Shahani, 1993; Green et al., 1993). Others have found differences in cue utilization which translated into differences in validities (Dougherty et al., 1986; Kinicki, Lockwood, Hom, & Griffeth, 1990; Zedeck et al., 1983). Also, differences in decision strategies have been related to differences in judged interviewer effectiveness (Graves & Karren, 1992). More skilled interviewers have elicited more information and made more accurate judgments (Motowidlo et al., 1992), and interviewers higher in conscientiousness and other positive personality traits have made more accurate judgments (Pulakos, Nee, & Kolmstetter, 1995). However, recent evidence suggests that differences in interviewer validities may be due to sampling error (Pulakos, Schmitt, Whitney, & Smith, 1996), but that evidence was based on an interview that was structured on other components which may have diminished the possible effects of interviewer differences.

The range of structure is from one person conducting all interviews to different people conducting each interview. Using one interviewer is often impractical, so the level of structure is a matter of degree, with relatively fewer interviewers considered more structured. A compromise with panel interviews is to keep one member the same (Handyside & Duncan, 1954; Landy, 1976). Unfortunately, few articles mention this component of structure; thus, the range of typical practice cannot be ascertained.

Effects on Reliability, Validity, and User Reactions. Certainly, using the same interviewers should increase test-retest reliability (Table 2). This should influence the reliability of both the evaluation and the content. Error variance due to interactions with candidates should be reduced due to less variation in interviewers. Fewer interviewers should reduce contamination to the extent constant rating errors and differences in skill across interviewers are reduced. Conversely, using fewer interviewers might increase perceptions of EEO bias if it makes the process appear more idiosyncratic and subjective.

Research and Practice Issues. Although there are good arguments and some

evidence, there is not a clear understanding of differences in interviewer validities. This idiographic approach is an area for future research (Guion, 1987). Research might continue to emphasize strategies of cue utilization and interviewer behaviors to learn why such differences occur (Dougherty et al., 1986; Kinicki et al., 1990; Motowidlo et al., 1992).

If consistent differences exist, then research should focus on possible solutions. One solution is interviewing training (see component 14 below). Another is selecting better interviewers. No research has been conducted, but it may be possible to hypothesize distinguishing attributes between more versus less effective interviewers. For example, having experience with the job being staffed (e.g., incumbents and supervisors) might enhance knowledge of job requirements, as well as enhance candidate and interviewer reactions. Drawing from other literatures, Dipboye (1992) suggests these attributes might include verbal reasoning, intelligence, listening skills, self-monitoring, ability to decode nonverbal behavior, and motivation to be accurate.

Another possible solution is to use a high level of structure on other components so that different interviewers do not matter. For example, using a highly structured interview and only modestly skilled interviewers, M. Campion et al. (1994) found an interrater reliability of .97. This solution may be practical in situations where the number of candidates or other pragmatic concerns make it infeasible to use the same interviewers.

13. Do Not Discuss Candidates or Answers Between Interviews

Explanation and Alternatives. Discussing candidates may lead to irrelevant information entering the evaluation process, as well as instrumentation effects (Cook & Campbell, 1979) such as changing standards between interviews. This component easily applies to panels because the interviewers are all present and there is a temptation to discuss candidates. But it is just as applicable to individual interviewers when interviews are spread out in time because there are opportunities to discuss candidates with others outside the hiring process.

Most articles do not mention this component, so typical practice cannot be assessed. Nevertheless, two levels of structure can be envisioned, either

interviewers are instructed not to discuss candidates between interviews, or they are not so instructed. Only several experimental studies (Carlson et al., 1970; Heneman et al., 1975; Schwab & Heneman, 1969) and one field study (M. Campion et al., 1988) mention using this component of structure.

Effects on Reliability, Validity, and User Reactions. Reliability effects may be mixed. Avoiding conversations about candidates should enhance test-retest reliability because it reduces the potential for changing standards and other instrumentation effects (Table 2). However, avoiding conversations among interviewers might reduce interrater reliability and agreement because it prevents differences in evaluations from being identified and corrected.

Unless reliability is reduced, this component should enhance validity because it reduces the potential for contamination. It may enhance perceptions of procedural justice if it prevents consideration of irrelevant information and the emergence of implicit favorites (M. Campion et al., 1988). Conversely, this restriction of interviewer freedom may create negative reactions.

Research and Practice Issues. All predicted effects of this component need to be subjected to empirical investigation. The potential positive and negative effects suggest this component cannot be unequivocally recommended without further evidence. Also, there may be interaction effects with other components. In particular, this component may not matter if other components are structured, and the potential negative effects on interrater reliability and agreement may be prevented by proper training.

14. Provide Extensive Interviewing Training

Explanation and Alternatives. Training is probably the most common step taken by companies to improve their interview (Dipboye, 1992). However, training is less a component itself than it is a way to ensure other components are implemented correctly. Interviewing training is analogous to training test administrators, which is highly recommended by professional testing guidelines (Society for Industrial and Organizational Psychology, 1987). Components of interview structure can be easily taught (Howard, Dailey, & Gulanick, 1979). Table 5 presents the topical content and teaching processes used in

interviewing training programs based on a content analysis of the literature.

Interviews require more highly skilled administrators than other selection devices, such as tests. Thus, training has been discussed since the early literature (Anderson, 1954; Bolanovich, 1944; Handyside & Duncan, 1954; Hovland & Wonderlic, 1939; Raines & Rohrer, 1955; Wonderlic, 1942). Although many articles do not mention details of their training, most probably begin with a description of the background and purpose of the interview (Anderson, 1954; Bolanovich, 1944; Walters et al., 1993). Many programs then discuss the interview itself (Bolanovich, 1944; Carlson et al., 1971; Carrier et al., 1990; Hovland & Wonderlic, 1939; Motowidlo et al., 1992; Pulakos & Schmitt, 1995; Robertson et al., 1990; Walters et al., 1993). The training may include how to write interview questions (Janz, 1982; Latham et al., 1980; Orpen, 1985; Roth & J. Campion, 1992) or, more typically, how to use questions already written.

Training frequently includes a discussion of job requirements, so that interviewers understand how the questions are related to the job (Pursell et al., 1980). Some programs have trainees complete a job analysis survey (Green et al., 1993). Many also discuss rapport building techniques (Motowidlo et al., 1992; Roth & J. Campion, 1992; Robertson et al., 1990).

Training how to ask questions is important if there is discretion as to which questions to select from an array or how to probe (Carlson et al., 1971; Janz, 1982; Motowidlo et al., 1992; Orpen, 1985; Pulakos & Schmitt, 1995; Robertson et al., 1990; Wonderlic, 1942). A common topic is how to evaluate answers and use rating scales (Borman, 1982; Carrier et al., 1990; Green et al., 1993; Latham et al., 1980; Maurer & Fay, 1988; Motowidlo et al., 1992; Orpen, 1985; Pulakos & Schmitt, 1995; Pursell et al., 1980; Robertson et al., 1990; Vance et al., 1978; Walters et al., 1993). It is also common to discuss avoiding rating errors (Carrier et al., 1990; Maurer & Fay, 1988; Walters et al., 1993).

Notetaking is addressed in many programs, either to prepare for rating (Janz, 1982; Mayfield et al., 1980; Orpen, 1985; Robertson et al., 1990) or to build documentation (Pursell et al., 1980). EEO laws and requirements are

addressed in many programs (Carrier et al., 1990; Maurer & Fay, 1988; Roth & J. Campion, 1992). Finally, some programs deal with how hiring decisions should be made from interview results, such as weighting questions (Janz, 1982; Orpen, 1985) and using ranking or cut-off scores (Pursell et al., 1980).

There are also similarities in training process. Lecture and discussion are most common. Many programs use behavioral modeling to show correct interviewing procedures, role-playing and practice interviews to build skills, and then feedback and reinforcement from the trainer and classmates to improve the skills (M. Campion et al., 1994; Carlson et al., 1971; Dougherty et al., 1986; Green et al., 1993; Maurer & Fay, 1988; Motowidlo et al., 1992; Pulakos & Schmitt, 1995; Robertson et al., 1990; Roth & J. Campion, 1992; Walters et al., 1993). Even though these techniques are popular now, their value for interview training was recognized early (Wonderlic, 1942). Videotaping can be used for many reasons, including modeling proper behaviors (Dougherty et al., 1986; Robertson et al., 1990), presenting candidates to the trainees to evaluate (Heneman et al., 1975; Maurer & Fay, 1988; Vance et al., 1978), and recording interviews for feedback (Motowidlo et al., 1992; Roth & J. Campion, 1992).

Many programs provide interviewing manuals (Anderson, 1954; Carlson et al., 1971; Carrier et al., 1990; Grove, 1981, Holt, 1958; Hovland & Wonderlic, 1939; Mayfield et al., 1980; Raines & Rohrer, 1955; Reeb, 1969; Robertson et al., 1990; Shaw, 1952). It is likely that most classes are small and highly interactive. Finally, the programs range from several hours (Anderson, 1954; Raines & Rohrer, 1955) to a week (Bolanovich, 1944) but the majority are one or two days (Borman, 1982; Dougherty et al., 1986; Green et al., 1993; Handyside & Duncan, 1954; Maurer & Fay, 1988; Motowidlo et al., 1992; Pulakos & Schmitt, 1995; Robertson et al., 1990; Roth & J. Campion, 1992; Walters et al., 1993).

Unlike most other components, it is difficult to see gradations in the degree of structure. However, extensive training would appear to include most topics and processes in Table 5 and take a day or two.

Effects on Reliability, Validity, and User Reactions. Training could have many positive effects. If properly trained, interviewers should be able to

elicit interview content and evaluate it consistently, thus improving all forms of test-retest and interrater reliability and agreement (Table 2). Trained interviewers should be able to put candidates at ease, thus enhancing their consistency and minimizing differences in interactions. The only negative effect is on internal consistency because many programs encourage avoiding the appearance of "halo" effects (i.e., high correlations among ratings). The Conway et al. (1995) meta-analysis supports the prediction of a positive effect on interrater reliability and no effect on internal consistency.

Although many studies have shown that structured interviews with proper training can have high reliability, there is little research demonstrating the unique effects of training. Vance et al. (1978) found that brief "rating error" training did not improve reliability. However, this type of training tells raters to avoid the appearance of error, but may not improve reliability (Bernardin & Pence, 1980). This led Maurer and Fay (1988) to combine "frame of reference" training (Bernardin & Buckley, 1981), which emphasizes consistency of evaluation, with rater error training, but again found no effects. Neither study is a strong test because their primary focus on rating errors is only a small part of interviewing training (compared to Table 5).

Training focusing on job-related questions and objectively scoring answers, and away from non job-related questions and evaluation approaches, should not only increase job-relatedness, but should also decrease deficiency and contamination. The training by Vance et al. (1978) on rating errors had no positive effect on accuracy. However, Dougherty et al. (1986) found that more extensive training involving job-related questions, rating scales, and practice interviews with feedback did improve the interviewers' predictive validities. Likewise, Pulakos et al., (1995) found that an extensive interviewer training program did improve rating accuracy.

EEO is a key part of most training programs, emphasizing candidate legal rights, questions that might be discriminatory, and how bias can enter into interview decisions. Such training should reduce EEO bias, and reviews of court cases have found that training relates to verdicts for the organization

(Gollub-Williamson et al., 1995). Candidate reactions should be positive if interviewers are trained in establishing rapport and putting candidates at ease, as well as being organized for the interview. Interviewers may respond positively to training, not only because training programs often elicit positive reactions, but because such training might help them make better decisions (Latham & Finnegan, 1993).

Research and Practice Issues. A primary research issue is whether training has unique positive effects. Based on the popularity of training and the number of consulting firms offering such programs, it can be speculated that organizations are spending more money on interviewing training than on any other single personnel selection system. As is unfortunately the case with most areas of training, little money is spent on evaluation (Goldstein, 1991).

Another issue is whether there is a trade-off between training and other components. If an interview is highly structured in other respects, extensive training may not be needed because there would be fewer discretionary behaviors for the interviewer, less subjectivity, and fewer complex skill requirements. Both Vance et al. (1978) and Maurer and Fay (1988) compared training with other components of structure (e.g., behavioral rating scales and situational questions, respectively), and both found that only the other components had a positive effect. If there are a large number of interviewers, it may be more cost effective to develop a highly structured interview and a short training program than to provide extensive training on a less structured interview.

Another question is whether training candidates improves the interview (Dipboye, 1992). This training could be an orientation session or a handout focusing on such topics as what to expect during the interview, how to give good answers (e.g., specific examples of behavior), and how to prepare. Although training in interview-taking skills has uncertain benefits for experienced candidates (M. Campion & J. Campion, 1987), an explanation of the interview might help inexperienced candidates, and it may reduce anxiety and enhance reactions of all candidates.

A final issue is whether the effects of training decay over time. The

importance of following up on interviewers to ensure proper implementation was recognized early (Wonderlic, 1942). There have been cases where interviewers failed to follow structured interview processes. For example, Latham and Saari (1984) found that some interviewers did not score each answer as instructed, but instead used the scoring guide only to form overall impressions. Weekley and Gier (1987) found that, despite training, some interviewers read the rating scale anchors to the candidates and then asked which was best. Interviewers may add their "personal touch," and the variance in implementation may increase over time after training (Dipboye & Gaugler, 1993, p. 156).

One means of training interviewers and preventing decay of structure is to put the interview protocol on a computer (Green, 1995). The interviewer could read questions, type notes, and rate answers on the computer. Also, this would allow structured prompting and question branching, as well as a wide range of normative data, feedback, and other on-line analyses.

15. Use Statistical Rather than Clinical Prediction

Explanation and Alternatives. Different interviewers typically weigh information differently (Graves & Karren, 1992), thus another way to enhance structure is to use statistical procedures rather than interviewer judgments to combine information to make predictions (Dipboye, 1992). There was a controversy on statistical versus clinical prediction in the 1950s (Meehl, 1954). The debate was over whether purely statistical or mechanical methods of combining data were better than clinical methods where expert judgment was used to consider all the unique nuances of the data and situation in each case to render a more accurate prediction. Early interviewing research was embroiled in this controversy (Holt, 1958; Rafferty & Deemer, 1950; Trankell, 1959). The evidence from many contexts, however, overwhelmingly favors statistical prediction (Faust, 1984; Meehl, 1954, 1986; Sawyer, 1966).

The statistical versus clinical distinction applies to measurement as well as prediction (Sawyer, 1966). Information can be collected in mechanical ways with little judgment and subjectivity or in clinical ways with intuition and perceptions guiding the information collected. In the present paper, this was

addressed by the other components of structure. Using the same questions, rating scales, and so on, mechanizes the measurement process. The concern in this section is with statistical prediction based on the measurements.

There are three situations where statistical prediction is relevant in structured interviewing. First, multiple ratings from different questions or dimensions must be combined to make predictions. A clinical approach would combine ratings subjectively. No articles were found using this approach. A statistical approach would combine ratings using a formula. Some formulas use differential weights for each rating, based on either judgment (Adams & Smeltzer, 1936; Janz, 1982; Maas, 1965; Orpen, 1985) or empirical relationships with a criterion (Dougherty et al., 1986; Walters et al., 1993).

Other formulas use unit weighting (i.e., weights of 1.0) wherein each rating is given equal weight. Unit weighting does not require cross-validation because there is no capitalization on chance. It typically yields validities as high as differential weighting and is more robust (Einhorn & Hogarth, 1975; Wainer, 1976). This usually involves taking a simple average or sum across all questions or dimensions, and it is the most common approach (Arvey et al., 1987; M. Campion et al., 1988, 1994; Delery et al., 1994; DuBois & Watson, 1950; Freeman et al., 1942; Green et al., 1993; Hakel, 1971; Hovland & Wonderlic, 1939; Komives et al., 1984; Landy, 1976; Latham et al., 1980; Latham & Skarlicki, 1995; Maurer & Fay, 1988; Motowidlo et al., 1992; Pulakos & Schmitt, 1995; Reynolds, 1979; Robertson et al., 1990; Roth & J. Campion, 1992; Shahani et al., 1991; Stohr-Gillmore et al., 1990; Weekley & Gier, 1987).

Either statistical approach is more structured than the clinical approach. Probably unit weighting is preferred due to its simplicity, lack of need for criterion data or cross-validation, and likely comparable results.

The second situation is combining data across interviewers. The most structured approach is to use a formula, of which averaging or summing is most common (M. Campion et al., 1988, 1994; Delery et al., 1994; DuBois & Watson, 1950; Freeman et al., 1942; Glaser et al., 1958; Komives et al., 1984; Landy, 1976; Reynolds, 1979). A less structured and more clinical approach has

interviewers discuss differences to consensus. This approach is common (Flynn & Peterson, 1972; Green et al., 1993; Grove, 1981; Latham et al., 1980; Latham & Skarlicki, 1995; Pulakos & Schmitt, 1995; Roth & J. Campion, 1992; Stohr-Gillmore et al., 1990; Tarico et al., 1986), possibly because the discussion might lead to more accurate consensus ratings or because such discussions are typical in other assessment contexts (Thornton & Byham, 1982). The least-structured and purely clinical approach would have one person combine information on a case-by-case basis (Huse, 1962).

The third situation is outside the interview per se, but is considered here because it has a substantial impact on validity. In many contexts, the interview is combined with other information (e.g., test scores, references, etc.) to make final decisions. This does not refer to withholding ancillary information before the interview (component number 6), instead it refers to combining information after the interview. This creates slippage in the structure of the hiring process because it allows subjectivity. Judgments are often made clinically because the data are not directly comparable. One way to structure the process is to convert scores to percentiles or standard scores so they will be comparable and then develop weights (probably through judgment) to combine the data. The only requirement to make the prediction statistical is that the rules for combining data be consistently applied.

Effects on Reliability, Validity, and User Reactions. Based on the evidence favoring statistical prediction, it is expected that the reliability will improve (Table 2). Internal consistency may also improve because a mechanical formula for combining items should increase item-total correlations. A statistical approach should reduce the likelihood that contaminating information will enter the score or that information will be neglected (i.e., deficiency). Consistency in making interview decisions predicts favorable court decisions (Gollub-Williamson et al., 1995) and should, thus, reduce EEO bias. User reactions should be unaffected because statistical rules are largely invisible to candidates and interviewers. Also, interviewer reactions are hard to predict because consensus discussions may enhance commitment to the

final ratings, but discussing each rating can make the process more effortful.

It should be noted, however, that meta-analyses have yielded mixed results. Conway et al. (1995) found mechanical combination yielded higher reliabilities than subjective methods when multiple ratings were used. But in Wiesner and Cronshaw (1988), six studies of panel interviews using averaging had a mean validity of .23 (.41 corrected) compared to .35 (.64 corrected) for seven studies using consensus. The interviews using consensus also had higher reliabilities (.74 versus .84). These results were contrary to their hypothesis and past findings. In a recent individual study, Pulakos et al. (1996) found that consensus and averaged ratings had similar validities in a structured interview. Perhaps consensus ratings reflect the best of both approaches in that they require independent ratings considered in a consistent manner like the statistical approach, and rational resolution of differences like the clinical approach.

Research and Practice Issues. Most interviews combine ratings into total scores based on statistical rules. Averaging or summing is the most common, but sometimes differential weighting is desired due to differences in importance between job requirements. Unit weighting can still be used in these cases, with differences in requirements reflected in the number of questions written to assess each requirement (M. Campion et al., 1988).

The need for a consensus discussion when combining ratings across interviewers is an unresolved issue, however. This is a good topic for future research because current evidence is counter to usual findings that statistical methods are better, and few studies have been conducted. Future research should determine how a consensus discussion might add value, similar to research on the assessment center (Sackett & Wilson, 1982). For example, discussion might identify errors in perceptions, clarify misinterpretations, or confront biases. A compromise is to average across interviewers and only discuss large differences (e.g., 2 points or more on a 5-point scale; M. Campion et al., 1988) or simply drop deviant ratings (Freeman et al., 1942).

Combining the interview with information such as test scores, resumes,

references, etc., is an important issue for future research. Of the situations considered here, this may be the best way to improve the hiring process through statistical prediction. Some evidence supports the value of statistical over clinical methods in combining diverse information in assessment centers (Borman, 1982; Wollowick & McNamara, 1969), but more research is needed.

Discussion

Structure and Psychometric Properties

This review supports several conclusions. First, interview structure is more complex than previously supposed. Structure is often equated with merely using the same questions or rating scales. However, this review identified 15 discernable components. Second, the interview can be easily enhanced by using at least some of these components. All had either empirical or rational links to enhanced reliability or validity. With so many ideas and such a large body of supportive literature, there is little reason to continue to use traditional unstructured interviews. Third, the improvement of this popular technique should be a high priority for future research. The ubiquity of interviews in organizations, along with the numerous issues surrounding virtually every component, suggest a great potential payoff for research in this area.

There is no consensus among experts as to which components of structure are most important (M. Campion & Palmer, 1995). Research is needed that manipulates components of structure to determine which are most important and how much structure is minimally required. Nevertheless, based on (a) amount of previous research, (b) strongest psychometric arguments, and (c) greatest opportunity for gain over typical practices, some components appear more important. Regarding content, the use of job analysis, same questions, and better questions appear more important than other components. The others have less certain value (e.g., limiting prompting and controlling ancillary information), can be addressed in other ways (e.g., no questions from candidate), or may not be a problem in practice (e.g., longer interview). Regarding evaluation, rating each answer or having multiple scales, using anchored scales, and training appear more important. The other components have

less certain value (e.g., no discussion between interviews), are common already (e.g., notetaking, statistical prediction), or may not be a problem if other components are structured (e.g., multiple or same interviewers).

Structure and User Reactions

Most potential negative effects are in terms of user reactions. Components such as limited prompting and follow-up, longer interviews, control of ancillary information, and no questions from candidates may cause unfavorable reactions from both candidates and interviewers. Further, panel interviews may be stressful for candidates, and being required to take detailed notes and not allowed to discuss candidates may be resented by interviewers.

These possible drawbacks may not necessarily cause participants to dislike and resist the interview for several reasons. First, they are based primarily on a rational analysis of expected outcomes and not on empirical research. Second, the obvious fairness of structured interviews should be recognizable by both parties (M. Campion et al., 1988; Pursell et al., 1980), especially if properly explained. Third, users realize the value of such interviews for making difficult personnel decisions (Latham & Finnegan, 1993). Fourth, the clear validity advantages make them a compelling business choice, despite minor inconvenience for participants. Fifth, the success of consulting firms in marketing these interviews attests to their appeal to organizations, even though the most popular systems are not the most highly structured (Hauenstein, 1995). Finally, the components with the most negative expected user impact (e.g., limiting prompting, controlling ancillary information, not allowing questions from candidates, etc.) may be the least critical to the interview's psychometric performance. If a very high level of structure results in some negative reactions, a more moderate level might be acceptable.

Clearly, future research should not only examine which components create negative reactions, but also which characteristics of participants predispose them to react negatively. For example, candidates with greater impression management needs may object more to structure (Fletcher, 1989). Likewise, highly experienced interviewers may object more to structure since it can

reduce the interview to a mindless exercise (Langer, Blank, & Chanowitz, 1978).

Another aspect of user reactions is motivation to conduct the interview properly. This is key to implementation (Pulakos, 1995). Motivation arises in several ways. First, interviewers may prefer the flexibility of unstructured interviews (Dipboye, 1994) and, thus, resist or modify the interview (Dipboye & Gaugler, 1993). Second, structure may reduce the enriching characteristics (M. Campion, 1988) of the interviewing task (Dipboye, 1995), thus inadvertently degrading intrinsic motivation. Third, specific motivations can influence ratings. For example, hiring quotas or tight labor markets may lead to inflated ratings (Carlson et al., 1971; Webster, 1982), and marginal candidates who will become co-workers may receive inordinately low ratings (Eder, 1989). Finally, one way to maintain motivation is to require accountability. It has been shown to lead to more accurate assessments (Rozelle & Baxter, 1981), and it may be enhanced by simple monitoring (Pulakos, 1995).

Other aspects of interviews could be structured to improve user reactions as well as other outcomes, but have been inadequately studied to draw any conclusions. For example, physical environment factors such as discomfort (e.g., hard chair), lack of privacy (e.g., interview in public area), and lack of personal presence (e.g., telephone or computer interviews) may create negative reactions and reduce reliability by distracting the participants. Psychological environment factors such as rapport, explaining the interview, and stress may influence candidate reactions, candidate consistency, and interactions with interviewers. Finally, interviews are not usually structured to either avoid or measure nonverbal behavior, which can influence ratings (Forbes & Jackson, 1980; Motowidlo & Burnett, 1995).

Theory Relevant to Structured Interviews

Another conclusion is that theory has not played an important role in this area. Past research was very applied; it was conducted to solve practical problems rather than to test theory. This paper relied mainly on psychometric theory to explain the operation of structured interviews. However, other more content- (as opposed to measurement-) oriented theories may offer additional

insight. For example, cognitive theory (Lord & Maher, 1991) might be used to consider underlying mechanisms. Structure may reduce information processing requirements and potential for overload, thus allowing interviewers to attend more fully to candidate responses (Arvey, 1995). Structure may also clarify the cognitive schemata used to interpret responses (Green, 1995), thus allowing responses to be classified and judged more systematically and accurately.

Another potentially useful perspective is attribution theory (Kelley, 1967). Structure may reduce variance due to differences in attribution style. Tendencies to attribute candidate responses to either internal or external factors may be controlled with defined standards of evaluation. Similarly, interviewers may hold implicit personality theories (Schneider, 1973) for desirable candidate characteristics. Again, structure may reduce these differences by defining the important ones explicitly.

Finally, Webster (1982) describes several interviewer decision making models. A conflict model explains how conflict and stress influence decision making, an information processing model explains decision making in terms of mathematical models, and an affect model explains the role of feelings and preferences in decision making. Structure might define the decision making task such that the influence of these processes may be lessened.

The State of the Literature

Reviews of the literature often note the lack of detail in most articles. This review is no exception. Most studies did not contain enough information to judge the level of structure on all components. This is partly due to article length considerations and the fact that many studies focused on issues other than structure. Also, interviews are often just one of many selection procedures examined (e.g., within an assessment center) and may not have been the primary emphasis. Further progress in accumulating knowledge of interview structure would be enhanced if future studies reported fuller information.

More disconcerting is the overall quality of the literature. Much of it is old, clinical in orientation, conducted in ambiguous settings, or confounded in many ways. Studies tend to have small samples, simple criteria, restriction

of range, and measures with modest reliability and unknown construct validity.

These problems are troubling for meta-analyses. Such techniques can correct for statistical limitations (e.g., sample, range, and reliability), but they cannot make precise comparisons between components of structure when information is lacking, components are confounded, or sufficient primary studies not conducted.

An equally difficult issue is the unknown construct validity of many interviews. Interviews are measurement techniques that are not linked to particular constructs. If the content of interviews is unclear, meta-analytic results must be correspondingly ambiguous. To illustrate, meta-analyses have included clinical interviews. They differ from selection interviews in focus (i.e., maladjustment and psychopathology versus job performance) and time orientation (i.e., current identification versus future prediction). They also rely on complex clinical judgment that may not easily translate into practice for managers. Such studies should not be used in meta-analyses, or they should be analyzed separately (McDaniel et al., 1994). More attention should be given to what constructs are measured by interviews as well as how they are measured.

Conclusion

Structured interviews are clearly superior psychometrically. Yet, administrative innovations, such as structured interviews, are rarely based on technical merit (Johns, 1993). Instead, researchers might have to emphasize environmental threats (e.g., low candidate quality), government regulations (e.g., EEO laws), or simple imitative or competitive processes to convince organizations to adopt them (Johns, 1993).

In conclusion, the selection interview can be enhanced by using some of the many possible components of structure, and the improvement of this popular selection procedure should be a high priority for future research and practice.

References

Adams, C. R., & Smeltzer, C. H. (1936). The scientific construction of an interview chart. Personnel, 13, 14-19.

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

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

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

Armstrong, J. S., Denniston, W. B., Jr. & Gordon, M. M. (1975). The use of the decomposition principle in making judgments. Organizational Behavior and Human Performance, 14, 257-263.

Arvey, R. D., (1995, May). Panelist. In M. A. Campion & D. K. Palmer (Chairs), Taking stock of structure in the employment interview. Panel discussion presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

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

Arvey, R. D., & Faley, R. H. (1988). Fairness in selecting employees (2nd ed.). Reading, MA: Addison-Wesley.

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

Barrett, G. V., Svetlik, B., & Prien, E. P. (1967). Validity of the job-concept interview in an industrial setting. Journal of Applied Psychology, 51, 233-235.

Beatty, R. N. (1986). The five-minute interview. New York: Wiley.

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

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training:

Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

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 Education Psychology, 35, 545-553.

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

Butler, S. K., & Harvey, R. J. (1988). A comparison of holistic versus decomposed rating of position analysis questionnaire work dimensions. Personnel Psychology, 41, 761-771.

Campbell, J. T. (1962). Assessments of higher-level personnel: 1. Background and scope of the research. Personnel Psychology, 15, 57-62.

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

Campion, J. E., & Arvey, R. D. (1989). Unfair discrimination in the interview. In R. W. Eder & G. R. Ferris (Eds.), The employment interview: Theory, research, and practice (pp. 61-73). Newbury Park, CA: Sage.

Campion, M. A. (1988). Interdisciplinary approaches to job design: A constructive replication with extensions. Journal of Applied Psychology, 73, 467-481.

Campion, M. A. (1993). Article review checklist: A criterion checklist for reviewing research articles in applied psychology [Editorial]. Personnel Psychology, 46, 705-718.

Campion, M. A., & Campion, J. E. (1987). Evaluation of an interviewee skills training program in a natural field experiment. Personnel Psychology, 40, 675-691.

Campion, M. A., Campion, J. E., & Hudson, J. P., Jr. (1994). Structured interviewing: A note on incremental validity and alternative question types. Journal of Applied Psychology, 79, 998-1002.

Campion, M. A., & Palmer, D. K. (Chairs) (1995, May). Taking stock of structure in the employment interview. Panel discussion presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

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.

Carlson, R. E., Schwab, D. P., & Heneman, H. G., III. (1970). Agreement among styles of selection interviewing. Journal of Industrial Psychology, 5, 8-17.

Carlson, R. E., Thayer, P. W., Mayfield, E. C., & Peterson, D. A. (1971). Improvements in the selection interview. Personnel Journal, 50, 268-275, 317.

Carrier, C. A., & Titus, A. (1979). The effects of note-taking: A review of studies. Contemporary Educational Psychology, 4, 299-314.

Carrier, M. R., Dalessio, A. T., & Brown, S. H. (1990). Correspondence between estimates of content and criterion-related validity values. Personnel Psychology, 43, 85-100.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. Journal of Applied Psychology, 80, 565-579.

Cook, T. D., & Campbell, D. T. (1979). Quasi experimentations: Designs and analysis for field settings. Boston: Houghton Mifflin.

Cornelius, E. T., III., & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. Journal of Applied Psychology, 65, 155-163.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Dalessio, A. T., & Silverhart, T. A. (1994). Combining biodata test and interview information: Predicting decisions and performance criteria. Personnel Psychology, 47, 303-315.

Delery, J. E., Wright, P. M., McArthur, K., & Anderson, D. C. (1994). Cognitive ability tests and the situational interview: A test of incremental validity. International Journal of Selection and Assessment, 2, 53-58.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

DeNisi, A. S., Robbins, T., & Cafferty, T. P. (1989). Organization of information used for performance appraisals: Role of diary-keeping. Journal of Applied Psychology, 74, 124-129.

Dillman, D. A. (1978). Mail and telephone surveys: The total design method. New York: Wiley-Interscience.

Dipboye, R. L. (1992). Selection interviews: Process perspectives. Cincinnati, OH: South-Western.

Dipboye, R. L. (1994). Structured and unstructured selection interviews: Beyond the job-fit model. In G. R. Ferris (Ed.), Research in personnel and human resources management: Vol. 12 (pp. 79-123). Greenwich, CT: JAI Press.

Dipboye, R. L. (1995, May). Panelist. In M. A. Campion & D. K. Palmer (Chairs), Taking stock of structure in the employment interview. Panel discussion presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Dipboye, R. L., Fontenelle, G. A., & Garner, K. (1984). Effects of previewing the application on interview process and outcomes. Journal of Applied Psychology, 69, 118-128.

Dipboye, R. L., & Gaugler, B. B. (1993). Cognitive and behavioral processes in the selection interview. In N. Schmitt, W. C. Borman, & Associates (Eds.), Personal selection in organizations (pp. 135-170). San Francisco: Jossey-Bass.

Dipboye, R. L., Gaugler, B. B., & Hayes, T. (1990, April). Individual differences among interviewers in the incremental validity of their judgments. Paper presented at the 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.

Drake, J. D. (1982). Interviewing for managers: A complete guide to employment interviewing (Rev. ed.). New York: AMACOM.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 1 (2nd ed., pp. 577-636). Palo Alto, CA: Consulting Psychologist Press.

Dreher, G. F., Ash, R. A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. Personnel Psychology, 41, 315-327.

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

DuBois, P. H., & Watson, R. I. (1950). The selection of patrolmen. Journal of Applied Psychology, 34, 90-95.

Eder, R. W. (1989). Contextual effects on interview decisions. In R. W. Eder & G. R. Ferris (Eds.), The employment interview: Theory, research, and practice (pp. 113-126). Newbury Park, CA: Sage.

Edwards, J. C., Johnson, E. K., & Molidor, J. B. (1990). The interview in the admission process. Academic Medicine, 65, 167-177.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. Organizational Behavior and Human Performance, 7, 86-106.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 13, 171-192.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. Federal Register, 43, 38290-38315.

Faust, D. (1984). The limits of scientific reasoning. Minneapolis, MN: University of Minnesota Press.

Feild, H. S., & Gatewood, R. D. (1989). Development of a selection interview: A job content strategy. In R. W. Eder & G. R. Ferris (Eds.), The employment interview: Theory, research, and practice (pp. 145-157). Newbury Park, CA: Sage.

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.

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

Fletcher, C. (1989). Impression management in the selection interview. In R. A. Giacalone & P. Rosenfeld (Eds.), Impression management in the organization (pp. 269-281). Hillsdale, NJ: Earlbaum.

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

Forbes, R. J., & Jackson, P. R. (1980). Non-verbal behaviour and the outcome of selection interviews. Journal of Occupational Psychology, 53, 65-72.

ForsterLee, L., Horowitz, I. A., & Bourgeois, M. (1994). Effects of notetaking on verdicts and evidence processing in a civil trial. Law and Human Behavior, 18, 567-578.

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.

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

Gehrlein, T. M., Dipboye, R. L., & Shahani, C. (1993). Nontraditional validity calculations and differential interviewer experience: Implications for selection interviews. Educational and Psychological Measurement, 53, 457-469.

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

Glaser, R., Schwarz, 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.

Goldstein, I. L. (1991). Training in organizations. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 2 (2nd ed., pp. 507-619). Palo Alto, CA: Consulting Psychologists Press.

Gollub-Williamson, L. R., Campion, J. E., Malos, S. B., Campion, M. A., & Roehling, M. V. (1995). The employment interview on trial: Linking legal

defensibility with interview structure. Manuscript submitted for publication.

Graves, L. M., & Karren, R. J. (1992). Interviewer decision processes and effectiveness: An experimental policy-capturing investigation. Personnel Psychology, 45, 313-340.

Green, P. C. (1995, May). Panelist. In M. A. Campion & D. K. Palmer (Chairs), Taking stock of structure in the employment interview. Panel discussion presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Green, P. C., Alter, P., & Carr, A. F. (1993). Development of standard anchors for scoring generic past-behaviour questions in structured interviews. International Journal of Selection and Assessment, 1, 203-212.

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

Guion, R. M. (1987). Changing views for personnel selection research. Personnel Psychology, 40, 199-213.

Hakel, M. D. (1971). Similarity of post-interview trait rating intercorrelations as a contributor to interrater agreement in a structured employment interview. Journal of Applied Psychology, 55, 443-448.

Hakel, M. D. (1982). Employment interviewing. In K. M. Rowland & G. R. Ferris (Eds.), Personnel management (pp. 129-155). Boston: Allyn and Bacon.

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, J. G., Jr. (1972). Prediction of success on a distant Pacific island: Peace Corps style. Journal of Consulting and Clinical Psychology, 38, 181-190.

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

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 2 (2nd ed., pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.

Hauenstein, P. C. (1995, May). Panelist. In M. A. Campion & D. K.

Palmer (Chairs), Taking stock of structure in the employment interview. Panel discussion presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

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

Hickling, L. P., Hickling, E. J., Sison, G. F. P., Jr. & Radetsky, S. (1984). The effect of note-taking on a simulated clinical interview. Journal of Psychology, 116, 235-240.

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.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. Journal of Abnormal and Social Psychology, 56, 1-12.

Hough, L. M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. Journal of Applied Psychology, 69, 135-146.

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

Howard, G. S., Dailey, P. R., & Gulanick, N. A. (1979). The feasibility of informed pretests in attenuating response-shift bias. Applied Psychological Measurement, 3, 481-494.

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. Journal of Applied Psychology, 79, 184-190.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1995, May). Assessment of mental ability in the employment interview. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

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.

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

Johns, G. (1993). Constraints on the adoption of psychology-based personnel practices: Lessons from organizational innovation. Personnel Psychology, 46, 569-592.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), Nebraska Symposium on Motivation: Vol. 15 (pp. 192-238). Lincoln, NE: University of Nebraska Press.

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

Kesselman, G. A., & Lopez, F. E. (1979). The impact of job analysis on employment test validation for minority and nonminority accounting personnel. Personnel Psychology, 32, 91-108.

Kiewra, K. A., DuBois, N. F., Christian, D., McShane, A., Meyerhoffer, M., & Roskelley, D. (1991). Note-taking functions and techniques. Journal of Educational Psychology, 83, 240-245.

Kinicki, A. J., Lockwood, C. A., Hom, P. W., & Griffeth, R. W. (1990). Interviewer predictions of applicant qualifications and interviewer validity: Aggregate and individual analyses. Journal of Applied Psychology, 75, 477-486.

Kleiman, L. S., & Faley, R. H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. Personnel Psychology, 38, 803-833.

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.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Langdale, J. A., & Weitz, J. (1973). Estimating the influence of job

information on interviewer agreement. Journal of Applied Psychology, 57, 23-27.

Langer, E., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal attraction. Journal of Personality and Social Psychology, 36, 635-642.

Latham, G. P., & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler, J. L. Farr, & M. Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives (pp. 41-55). Hillsdale, NJ: Lawrence Erlbaum.

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.

Latham, G. P., & Skarlicki, D.P. (1995). Criterion-related validity of the situational and patterned behavior description interviews with organizational citizenship behavior. Human Performance, 8, 67-80.

Locke, E. A., & Latham, G. P. (1984). Goal-setting: A motivational technique that works. Englewood Cliffs, NJ: Prentice-Hall.

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

Lord, R. G., & Maher, K. J. (1991). Cognitive theory in industrial and organizational psychology. In M. D. Dunnette & L. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 2 (2nd ed., pp. 1-62). Palo Alto, CA: Consulting Psychologists Press.

Lowry, P. E. (1994). The structured interview: An alternative to the assessment center? Public Personnel Management, 23, 201-215.

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

Macan, T. H., & Dipboye, R. L. (1994). The effects of the application on processing of information from the employment interview. Journal of Applied Social Psychology, 24, 1291-1314.

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the employment

interview: A meta-analysis. International Journal of Selection and Assessment, 1, 18-26.

Maurer, S. D., & Fay, C. (1988). Effect of situational interviews, conventional structured interviews, and training on interview rating agreement: An experimental analysis. Personnel Psychology, 41, 329-344.

Mayfield, E. C. (1964). The selection interview: A reevaluation of published research. Personnel Psychology, 41, 239-260.

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

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 651-696). Chicago: Rand McNally.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. Journal of Applied Psychology, 79, 599-616.

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

Meehl, P. E. (1954). Clinical versus statistical prediction. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. Journal of Personality Assessment, 50, 370-375.

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

Miller, M. J. (1992). Effects of note-taking on perceived counselor social influence during a career counseling session. Journal of Counseling Psychology, 39, 317-320.

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

Morse, M., & Hawthorne, J. W. (1946). Some notes on oral examinations. Public Personnel Review, 7, 15-18.

Motowidlo, S. J., & Burnett, J. R. (1995). Aural and visual sources of validity in structured employment interviews. Organizational Behavior and Human Decision Processes, 61, 239-249.

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

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 3 (2nd ed., pp. 61-138). Palo Alto, CA: Consulting Psychologists Press.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational, and goal-based perspectives. Thousand Oaks, CA: Sage.

Nevo, B., & Berman, J. A. (1994). The two-step selection interview: Combining standardization with depth. Research & Practice in Human Resource Management, 2, 89-96.

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

Oskamp, S. (1965). Overconfidence in case-study judgments. Journal of Consulting Psychology, 29, 261-265.

Otis, J. L. (1944). Improvement of employment interviewing. Journal of Consulting Psychology, 8, 64-69.

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

Pulakos, E. D. (1995, May). Panelist. In M. A. Campion & D. K. Palmer (Chairs), Taking stock of structure in the employment interview. Panel discussion presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Pulakos, E. D., Nee, M. T., & Kolmstetter, E. B. (1995, May). Effects of training and individual differences on interviewer rating accuracy. In E. B.

Kolmstetter (Chair), Interviewer and contextual factors that make a difference in interview validity. Symposium presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. Personnel Psychology, 48, 289-308.

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. Personnel Psychology, 49.

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.

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

Putney, R. W. (1947). Validity of the placement interview. Personnel Journal, 26, 144-145.

Rafferty, J. A., & Deemer, W. L., Jr. (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: I. Consistency and variability in interview impressions of different psychiatrists. American Journal of Psychiatry, 111, 721-733.

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

Reynolds, A. H. (1979). The reliability of a scored oral interview for police officers. Public Personnel Management, 8, 324-328.

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. (1992). An analysis of the predictive power of the panel interview and pre-employment tests. Journal of Occupational and Organizational Psychology, 65, 51-60.

Rozelle, R. M., & Baxter, J. C. (1981). Influence of role pressures on the

perceiver: Judgments of videotaped interviews varying judge accountability and responsibility. Journal of Applied Psychology, 66, 437-441.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67, 401-410.

Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. Journal of Applied Psychology, 67, 10-17.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66, 178-200.

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

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. Personnel Psychology, 39, 91-108.

Schneider, D. J. (1973). Implicit personality theory: A review. Psychological Bulletin, 79, 294-309.

Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. Applied Psychological Measurement, 13, 19-32.

Schuh, A. J. (1980). Verbal listening skill in the interview and personal characteristics of the listeners. Bulletin of the Psychonomic Society, 15, 125-127.

Schuler, H. (1989). Construct validity of a multimodal employment interview. In B. J. Fallon, H. P. Pfister, & J. Brebner (Eds.), Advances in industrial organizational psychology (pp. 343-354). Amsterdam: North-Holland.

Schwab, D. P., & Heneman, H. G., III. (1969). Relationship between interview structure and interviewer reliability in an employment situation. Journal of Applied Psychology, 53, 214-217.

Scott, W. D. (1915, October). The scientific selection of salesmen. Advertising and Selling, 5-6, 94-96.

Shahani, C., Dipboye, R. L., & Gehrlein, T. M. (1991). The incremental contribution of an interview to college admissions. Educational and Psychological Measurement, 51, 1049-1061.

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

Smeltzer, C. H., & Adams, C. R. (1936). A comparison of graphic and narrative interview reports. Personnel, 13, 41-45.

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

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. (1993). Applicant reactions to selection procedure. Personnel Psychology, 46, 49-76.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the validation and use of personnel selection procedures (3rd ed.). College Park, MD: Author.

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. Journal of Personality and Social Psychology, 53, 81-93.

Steiner, I. D. (1972). Group process and productivity. NY: Academic Press.

Stewart, V., & Stewart, A. (with Fonda, N.) (1981). Business applications of repertory grids. London: McGraw-Hill.

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(2), 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.

Thornton, G. C., III. & Byham, W. C. (1982). Assessment centers and managerial performance. New York: Academic Press.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological

Review, 34, 273-286.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22, 358-376.

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

Tucker, D. H., & Rowe, P. M. (1977). Consulting the application form prior to the interview: An essential step in the selection process. Journal of Applied Psychology, 62, 283-287.

Tullar, W. L., Mullins, T. W., & Caldwell, S. A. (1979). Effects of interview length and applicant quality on interview decision time. Journal of Applied Psychology, 64, 669-674.

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

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

Vance, R. J., Kuhnert, K. W., & Farr, J. L. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. Organizational Behavior and Human Performance, 22, 279-294.

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.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 83, 213-217.

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

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

Wanous, J. P. (1980). Organizational entry. Reading, MA: Addison-Wesley.

Webster, E. C. (1959). Decision making in the employment interview. Personnel Administration, 22(3), 15-22.

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

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.

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

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

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. Journal of Occupational Psychology, 61, 275-290.

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

Wollowick, H. B., & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. Journal of Applied Psychology, 53, 348-352.

Wonderlic, E. F. (1942). Improving interview techniques. Personnel, 18, 232-238.

Wright, O. R., Jr. (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.

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

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

Table 1

Components of Interview Structure-----
Content

1. Base questions on a job analysis
2. Ask exact same questions of each candidate
3. Limit prompting, follow-up questioning, and elaboration on questions
4. Use better types of questions (e.g., situational, past behavior, background, job knowledge, etc.)
5. Use longer interview or larger number of questions
6. Withhold or control ancillary information (e.g., application forms, test scores, etc.)
7. Do not allow questions from candidate until after the interview

Evaluation

8. Rate each answer or use multiple scales
 9. Use detailed anchored rating scales
 10. Take detailed notes
 11. Use multiple interviewers
 12. Use same interviewer(s) across all candidates
 13. Do not discuss candidates or answers between interviews
 14. Provide extensive interviewing training
 15. Use statistical rather than clinical prediction (e.g., add up item scores rather than subjectively combine, average across interviewers rather than discuss to consensus, etc.)
-

Table 2

Effects of Interview Structure on Reliability, Validity, and User Reactions

Table 3

Examples of Different Types of Structured Interview Questions

Situational Questions:

1. Suppose a co-worker was not following standard work procedures. The co-worker was more experienced than you and claimed the new procedure was better. Would you use the new procedure?

2. Suppose you were giving a sales presentation and a difficult technical question arose that you could not answer. What would you do?

Past Behavior Questions:

3. Based on your past work experience, what is the most significant action you have ever taken to help out a co-worker?

4. Can you provide an example of a specific instance where you developed a sales presentation that was highly effective?

Background Questions:

5. What work experiences, training, or other qualifications do you have for working in a teamwork environment?

6. What experience have you had with direct point-of-purchase sales?

Job Knowledge Questions:

7. What steps would you follow to conduct a brainstorming session with a group of employees on safety?

8. What factors should you consider when developing a television advertising campaign?

Note. So that direct comparisons can be made, an example is presented to assess both teamwork (1, 3, 5, and 7) and sales attributes (2, 4, 6, and 8) for each type of question.

Table 4

Alternative Approaches for Anchoring Structured Interview Rating Scales

Question: Setting priorities and planning are important job requirements.

Can you please give specific examples from your past jobs or other experiences where you had to set priorities and plan your work?

<u>Scaled examples</u>	<u>Descriptions</u>	<u>Evaluations</u>	<u>Comparisons</u>
5. * On aircraft engines, I classified tasks into A/B/C system.	* Used a specific system involving listing the tasks and assigning priorities.	* Excellent answer	* Top 20% of candidates
4. * In food preparation, I found out what was needed and when, then made out a schedule.	* Considered task importance and did the most important first.	* Good answer	* Next 20% of candidates
3. * I did important tasks first.			* Middle 20% of candidates
2. * I did first come first			* Next 20% of candidates
1. serve, or I asked the supervisor.	* No real system used.	* Marginal answer	* Bottom 20% of candidates

Note. Illustration of four approaches to anchoring using a past behavior question assessing the ability to work without supervision.

Table 5

Typical Structured Interviewing Training Programs

Content:

1. Background information and purpose of program
2. Nature of interview (preparation for interview)
3. Job requirements
4. Establishing rapport
5. Asking questions (and probing)
6. Evaluating answers
7. Using rating scales
8. Avoiding rating errors
9. Taking notes and documentation
10. Equal employment opportunity
11. Making hiring decisions

Process:

1. Lecture
 2. Discussion
 3. Modeling
 5. Practice and role-playing
 6. Feedback
 7. Videotaping
 8. Instructional materials
 9. Small classes and highly interactive
 10. One or two days in duration
-