PERSONNEL PSYCHOLOGY
1997, 50

A REVIEW OF STRUCTURE IN THE SELECTION INTERVIEW

MICHAEL A. CAMPION, DAVID K. PALMER
Purdue University

JAMES E. CAMPION
University of Houston

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. Fifteen 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.

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

All narrative reviews have supported the use of structured interviews (Arvey & J. Campion, 1982; Harris, 1989; Mayfield, 1964; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949; Wright, 1969). As early as 1949, Wagner stated that all interviews should be conducted according to a

Special thanks to Robert L. Dipboye, Paul C. Green, J. Peter Hudson, Allen I. Huffcutt, Robert A. Jako, Gary P. Latham, Stephan J. Motowidlo, Elaine D. Pulakos, the graduate seminar on interviewing at the University of Houston (1995), and three anonymous reviewers for their comments on earlier drafts of this paper.

Correspondence and requests for reprints should be addressed to Michael A. Campion, Krannert School of Management, Purdue University, West Lafayette, IN 47907-1310.

Milton D. Hakel was the acting editor for this manuscript.

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. Across studies, corrected validities for unstructured interviews ranged from .14 to .33 and for structured interviews from .35 to .62 (Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989; see also Conway, Jako, & Goodman, 1995; Marchese & Muchinsky, 1993).

Overview of Paper

The purpose of this paper is to review the literature in order to summarize, integrate, and evaluate the many ways interviews can be structured. It 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.

The 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.

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 is intended to increase psychometric properties by increasing standardization or otherwise 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 components are divided into two categories: those that influence the *content* of the interview, or the nature of the information elicited, and those that influence the *evaluation* process, or the judgment of the information elicited. Although some components could be in both categories, the distinction is useful because it highlights a primary difference.

TABLE 1

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

	Reliability						Validity			User reactions		
	Test-retest	Inter-rater	Candid. consist.	Inter-cand. interaction	Internal consist.	Interrater agreement	Job-relatedness	Reduced deficiency	Reduced contam.	Reduced EEO bias	Candidate reactions	Interviewer reactions
Content												
1. Job analysis							+	+	+	+	+	+
2. Same questions	+	+	+	+				+	+	+		
3. Limit prompting	+	+	+	+				−	+	+	−	−
4. Better questions			+		+		+		+	+		+
5. Longer interview	+	+			+			+			−	−
6. Control ancillary information	+	+						−	+	+	−	−
7. No questions from candidate	+	+	+	+				−	+		−	−
Evaluation												
8. Rate each answer or use multiple scales	+	+			+			+	+			
9. Anchored rating scales	+	+				+	+	+	+	+		+
10. Detailed notes	+	+				+	+	+	+	+		−
11. Multiple interviewers	+	+		+	+	+		+	+	+	−	
12. Same interviewer(s)	+			+					+	−		
13. No discussion between interviews	+	−				−			+	+		−
14. Training	+	+	+	+	−	+	+	+	+	+	+	+
15. Statistical prediction	+	+			+			+	+	+		

Note: "+" means positive effect and "−" means negative effect.

The impact of each component is evaluated in terms of reliability, validity, and user reactions (Table 1). This evaluation is based on prior research, if available, or else on psychometric theory or logic. Six types of reliability are considered:

1. Test-retest reliability: Is the same interview content elicited, and is the evaluation process consistent each time by the interviewer?

2. Interrater reliability: Do different interviewers elicit the same content and evaluate candidates consistently?

3. 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?

4. Interviewer-candidate interaction: Does the interview limit error variance due to differences in interactions between interviewers and candidates? Would differences in personalities, communication styles, interpersonal attraction, or similar factors influence the results?

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

6. Interrater agreement: Do interviewers agree on their judgments, thus making similar decisions? Reliability refers to covariation; it is not the same as agreement, which refers to mean differences (Tinsley & Weiss, 1975).

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 reduce potential bias against subgroups of candidates protected by equal employment opportunity (EEO) laws? This includes adverse impact and disparate treatment, as well as perceptions of fairness. Some structured interviews were specifically developed to enhance legal defensibility (Pursell, M. Campion, & Gaylord, 1980). Research has found many components empirically related to court verdicts (J. Campion & Arvey, 1989; Gollub-Williamson, J. Campion, Malos, M. Campion, & Roehling, 1996).

2. Candidate reactions: Will candidates view the interview positively? Such reactions reflect the perceived (face) validity of selection procedures, and they influence job choice, affective reactions, and referrals. Candidates may prefer interviews over tests (Wilson, 1948), but

structured interviews may be viewed less positively (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 of structured interviews (Latham & Finnegan, 1993), but also the political value of unstructured interviews (Dipboye, 1994).

Review of Components of Structure

1. Base Questions on a Job Analysis

Explanation and alternatives. A variety of job analysis methods can be used to develop structured interviews, but critical incidents is the most common (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). Meetings with job experts are often used to collect incidents but surveys can be used (Weekley & Gier, 1987).

Critical incidents provide ideas for interesting and job-related questions. However, the development of questions from incidents is part of the art (or unwritten aspects) of structured interviewing. "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 incidents that best represent the dimensions are turned into questions (Latham et al., 1980), thus enhancing content validity.

Analogous to critical incidents analysis, contrasting groups of high and low performing employees are sometimes examined (Holt, 1958; Robertson et al., 1990). In repertory grid analysis (Stewart & Stewart, 1981), experts 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). 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).

Job experts, such as managers and interviewers, often write the questions (Janz, 1982; Latham et al., 1980; Orpen, 1985; Roth & J. Campion, 1992). Most articles do not discuss this detail, implying they are written by researchers.

There are two levels of structure on this component, either a job analysis can be conducted or there are at least three (unstructured) alternatives. First, many interviews are conducted by psychologists focusing on personality traits but are not based on job analysis (Bobbitt & Newman, 1944; Fisher, Epstein, & Harris, 1967; 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?). Third, an intuitive approach is used wherein interviewers ask whatever questions thought to be 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).

Job analysis is not expected to enhance reliability, but there might be a weak positive relationship if it limits the domain in the interview. The Conway et al. (1995) meta-analysis showed a low positive relationship between job analysis and reliability, which they interpreted as an indirect effect. It is expected to influence all three types of validity, however (Table 1). Job analysis should enhance job-relatedness, 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 a common frame of reference, interviewers might base the interview on idiosyncratic beliefs about job requirements (Dipboye, 1994).

Narrative reviews have emphasized the importance of job analysis for improving validity (Arvey & J. Campion, 1982; Harris, 1989; Schmitt, 1976). The meta-analytic review of Wiesner and Cronshaw (1988) found a corrected validity of .87 when a formal job analysis was conducted, .59 when an informal analysis was conducted, and .56 when it was unknown. McDaniel et al. (1994) found corrected validities of .50 and .39 for interviews based on a job analysis and .29 for psychological interviews not based on a job analysis. Also, 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 1). 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 support this assertion from court cases on interviewing (Gollub-Williamson et al., 1996). Job-relatedness should enhance face validity. Involving interviewers in the analysis should improve their acceptance.

Research and practice issues. If job analysis is more likely to identify knowledges, skills, and abilities, than personality traits and other attributes (Harvey, 1991), then a key question is whether structured interviews are more valid than unstructured interviews because they tap into cognitive ability? This is important because ability tests 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 (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 found a corrected correlation of .40 between interviews and ability tests (Huffcutt, Roth, & McDaniel, 1995). Interviews with higher cognitive loading were also more valid.

A related question is whether structured interviews have incremental validity beyond cognitive ability tests. Again, 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 to the cognitive loading of the interview; M. Campion et al. (1994) had the highest loading, and Pulakos and Schmitt had the lowest.

Future research could address this issue in several ways. First, studies should include measures of cognitive abilities so data can be accumulated. Second, constructs assessed by interviews should be examined at conceptual and empirical levels. 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

structure in the interview and the range of abilities in the sample should be considered when interpreting results.

Other research and practice topics evolve around the form of job analysis used for developing structured interviews. For example, how is information turned into interview questions, can unique insight be gained using contrasting groups or the repertory grid, and might behavioral consistency 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). Other methods of job analysis have not been examined for developing questions (e.g., protocol analyses, diaries, 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). Early authors suggested that the idea of structuring interviews for employment was inspired by Binet's success using structured interviews for intelligence testing (Wagner, 1949; Wonderlic, 1942).

Four levels of structure emerge on this component. These levels are similar to those used by Huffcutt and Arthur (1994), except prompting and follow-up questioning are considered separately here (in 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; Stohr-Gillmore, Stohr-Gillmore, & Kistler, 1990; Walters et al., 1993; Weekley & Gier, 1987). Using the same paralanguage would enhance structure further.

The second highest level requires primarily 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 initial questions (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 questions. Instead, these interviews only provide outlines of topics to cover (Ghiselli, 1966; Yonge, 1956), lists of desirable 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).

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).

It may be advantageous to sequence or group the questions. 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 preemployment tests and, as such, should be standardized samples of behavior. "Standardization implies uniformity of procedure in administration and scoring" (Anastasi, 1976, p. 25). Using the same questions may be the most basic way to convert the interview from a conversation into a scientific measurement.

This component may increase interrater and test-retest reliability (Table 1) because different interviewers will ask the same questions, and interviewers will ask the same questions across candidates. There may be less opportunity for interviewer-candidate interactions with the same questions, and candidate consistency may increase because questioning will follow a predictable pattern.

Asking questions in the same order might enhance reliability because it increases the consistency of the interview. Grouping questions has been shown to increase internal consistency (Schriesheim, Solomon, & Kopelman, 1989), but this is not shown in Table 1 because it is not due to using the same questions.

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. It may reduce contamination by preventing discussion of tangential topics and other biasing influences (Dipboye & Gaugler, 1993) and by reducing cognitive overload of the interviewer by focusing attention 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 1). It limits many sources of bias in the interview (Dipboye, 1994). This component is viewed as defensible by attorneys (Latham & Finnegan, 1993), and it relates positively to court cases on the interview (Gollub-Williamson et al., 1996).

This component may show both positive and negative effects on reactions, however. Candidates may view structured interviews as more face valid (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993), but prefer the freedom of unstructured interviews so they can display their credentials positively (Latham & Finnegan, 1993). A predetermined question order may inconvenience candidates prepared to describe their background in a different order.

Likewise, interviewers may prefer the freedom of unstructured interviews in order to maintain control over decision making, communicate organizational values, and qualitatively assess fit (Dipboye, 1994). Interviewers may also prefer a prepared set of questions to enhance face validity, help organize 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). Thus, user reactions are ambiguous.

Research and practice issues. One question is what level of structure is needed to attain high validity. Huffcutt and Arthur's (1994) meta-analysis indicates that structure beyond the second highest level is not needed. But meta-analyses are limited by the number and variety of studies available, and components tend to be confounded. Huffcutt and Arthur noted that there was more variance in validities at higher levels of structure, suggesting other potential moderators not assessed. For example, they noted that interviewers may have been highly trained, which could compensate for the reduced structure on this component. Their study did not control other components of structure, and their results contradict 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 may vary by the questions selected. Future research might examine optimal strategies for selecting items during an interview. Potentially, concepts from item response theory (Drasgow & Hulin, 1990) could be used wherein more difficult subsequent questions are asked if the candidate correctly answers previous questions, and vice versa. The goal is to get an accurate measure with the fewest number of questions.

Research is needed on user reactions. Do standardized questions bother candidates or instead enhance perceptions of fairness and preparation? Research is needed to determine if interviewers see this component as helpful (Latham & Finnegan, 1993) or as a restriction of their freedom (Dipboye, 1994). Past research on reactions has not been methodologically strong. Often interviews are simply described to respondents and their reactions measured, rather than studying reactions in real selection contexts.

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).

There are four levels of this component. The first and highest level is the prohibition of any prompting, follow-up questioning, or elaboration. Questions can be repeated, or candidates can be given the question on a card (Green et al., 1993). This level tends to occur with those approaches that 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 allow only limited or pre-planned prompts and follow-ups. As examples, suggested probes can be provided (Carlson et al. 1971; Mayfield et al., 1980), only probes can be used with the same wording as the assessment dimension, interviewers can be allowed a specified number of follow-up questions (Freeman et al., 1942), or interviewers can be allowed to ask, "Is there anything else you would like to add?" (Robertson et al., 1990).

The third level is to allow and encourage unlimited probes and follow-ups. Sometimes follow-ups 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 (Green et al., 1993; Hakel, 1971; Motowidlo et al., 1992; Reynolds, 1979). In panel interviews, probing can be by those not asking questions (Roth & J. Campion, 1992). Unlimited probing distinguishes semistructured from structured interviews (Heneman, Schwab, Huett, & Ford, 1975; Schwab & Heneman, 1969).

The fourth and lowest level is no guidance on probing and follow-up. This is the lowest level because some may probe and others may not. It is the most frequent level of structure on this component.

Effects on reliability, validity, and user reactions. Structuring may increase interrater reliability by decreasing variation between interviewers on the types and extent of prompts used (Table 1). Test-retest reliability may increase, and interviewer-candidate interactions decrease, because interviewers will use the same prompts and follow-ups across candidates. Finally, candidate consistency might increase because questioning will be less spontaneous.

This component may have varied effects on validity (Table 1). Prompting is intended to clarify answers and seek information, so limiting its use might create deficiency (Huffcutt & Arthur, 1994). However, it could create contamination if extraneous information is introduced or if candidates are coached into giving the right answers. The net effect on validity is unclear.

The effects on user reactions are also varied (Table 1). It may have a positive effect on reducing EEO bias if the potential for illegal questions or showing favoritism through coaching is reduced. Conversely, candidates may prefer prompting and follow-ups because they enhance the conversational nature of the interview, thus allowing the flexibility to 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-ups, and elaboration lead to increased or decreased validity? 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.

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 (e.g., "Could you explain further?" "Please provide an example.") would not change the construct, yet could be standardized (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 user reactions. Do these limitations make the interview too unnatural? Do they limit interviewer freedom unnecessarily? Would these limitations be accepted if their purposes 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.

Different types of questions cannot be neatly ordered in terms of structure. However, certain types of questions are more structured because they are more frequently used with high levels of structure on other components or because of their more focused nature.

One type that has been widely studied and is 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 candidates are asked what they would do. They are usually used with high structure on other components, and their highly specific nature 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). Past behavior questions are usually used with moderately high structure on other components, and their highly specific nature enhances their structure.

A third fairly structured type is background questions. They typically focus on work experience, education, and other qualifications (Carlson et al., 1971; Lopez, 1966; Mayfield et al., 1980; Roth & J. Campion, 1992).

TABLE 2

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.

A fourth relatively structured type is job knowledge questions. Interviews in the literature have tended to mix these questions in with other types (Arvey et al., 1987; M. Campion et al., 1988; Walters et al., 1993). They may ask candidates to describe or document their job knowledge, or they may ask candidates to demonstrate job knowledge. The highly specific nature of these questions enhances structure.

These four types of questions are illustrated in Table 2. Parallel examples are given to allow for comparison.

Other question types have been included in 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; M. Campion et al., 1988). No research has focused exclusively on these types, however.

Many question types are less structured. 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 avoid revealing weaknesses. They 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 1), but these expectations are speculative.

The most likely effect is enhanced validity (Table 1). 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 relationship between goals or intentions and future behavior (Locke & Latham, 1984). Past behavior questions may predict because of the axiom that past behavior predicts future behavior. Past behavior questions in the form of biodata have shown good validity (Mumford & Stokes, 1992). Most articles on situational and past behavior questions reviewed above present validity evidence. Job knowledge questions are supported by validity generalization research (Hunter & 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 improve user reactions. Reduced EEO bias is expected due to test fairness evidence for situational (M. Campion et al., 1988) and past behavior questions (Pulakos & Schmitt, 1995). Interviewer reactions may be positive if the questions allow better decisions (Latham & Finnegan, 1993).

Research and practice issues. The search for better question types has been popular recently. For example, some studies have compared questions. M. Campion et al. (1994) compared future (situational) versus past (behavior) questions in a sample of pulp mill employees. Both types had validity, but past questions were more valid. In a comparison in a law enforcement sample, Pulakos and Schmitt (1995) found that only past behavior questions were valid.

Conversely, Latham and Saari (1984) compared these questions among clerical personnel and found only situational questions were valid. Maurer and Fay (1988) similarly found that situational questions had higher interrater agreement in a laboratory study. Both studies operationalized past behavior questions in terms of broad inquiries about past experiences and training, rather than specific questions requiring

specific examples of past behavior. Latham and Skarlicki (1995) operationalized past behavior questions correctly in a college faculty sample and again found only situational questions valid.

Furthermore, McDaniel et al. (1994) compared 16 validity coefficients for situational with 127 coefficients for job-related interviews and found the former more valid. However, the job-related interviews were very heterogeneous, including 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 current evidence. Future research might compare questions on other dimensions. For example, situational questions might work better with candidates lacking sufficient work experience to respond to past behavior questions. Previous studies have used mainly experienced candidates. Also, past behavior questions may be less fakable due to their verifiable nature, but this has not been tested. Finally, future research should explore other question types, such as simulations and willingness questions, and future research should control structure on other components.

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.

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 2.

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 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 explained the reasoning behind their choice of interview length.

Effects on reliability, validity, and user reactions. The most direct effect of length is on internal consistency (Table 1). Length may indirectly

affect test-retest and interrater reliability because longer interviews may produce more stable and comparable measurements. 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 ($p < .10$) and did not control for other components of structure. Shorter interviews may have been more structured in this sample. Marchese and Muchinsky speculate that overly long interviews may collect too much information such that overload occurs, thus reducing decision quality (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 threat to structure is the uncontrolled use of ancillary information. This includes application forms, resumes, test scores, recommendations, previous interviews, transcripts, and so forth. Many interviews have allowed 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).

Two problems are created by ancillary information. First, it confounds the interpretation of the value of the interview. Validity may be due to the interview or this other information (Ulrich & Trumbo, 1965; Webster, 1959). Second, it creates unreliability if not available for all candidates or given to all interviewers, or if interviewers evaluate the information differently.

To enhance structure, ancillary information can be withheld (M. Campion et al., 1988; Latham et al., 1980). It can still be considered, but 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 or standardizing this information should increase test-retest and interrater reliability (Table 1). There is some evidence to this effect (Dipboye et al., 1984). The effects on validity are uncertain. Withholding information may reduce contamination if it is not valid or used incorrectly, but withholding it may increase deficiency if it is valid. The McDaniel et al. (1994) meta-analysis found validities were higher when test information was not available.

This result seems counter to the fact that such information could make the interview more complete (Tucker & Rowe, 1977) and some ancillary information, such as tests, have substantial validity (Hunter & Hunter, 1984). Conversely, test information can be used unreliably, prescreening on ancillary information can cause range restriction, and it may depend on the information (Dalessio & Silverhart, 1994). Also, most studies are ambiguous regarding the availability of information, thus making meta-analyses tenuous. Yet, the 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 information unrelated to the job. However, interviewers may react negatively to not having access to relevant information, and candidates may react negatively if interviewers are unaware of relevant information that was submitted before the interview.

Research and practice issues. Future research should examine the expected effects on reliability and validity and how the information can be standardized. A practical issue is how to avoid potential negative reactions. It may be enough to explain to users. Another practical issue is how to make sure critical background information is complete 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, yet uncontrolled questions from candidates reduce standardization by changing the interview content in unpredictable ways. Also, the relational control or dominance of the interview can be affected (Tullar, 1989). Thus, structure can be enhanced

by *not* allowing questions during the interview. Instead, time can be allowed outside the interview.

Most articles do not mention this component. Unstructured interviews are conversational, with both parties asking questions (Drucker, 1957), and some interviews are intended to provide information to candidates (Komives, Weiss, & Rosa, 1984). Only one article used this component; it allowed candidate questions in a later, nonevaluation interview (M. Campion et al., 1988).

Effects on reliability, validity, and user reactions. Not allowing questions from candidates should standardize the content, thus increasing test-retest and interrater reliability (Table 1). Reducing the conversational nature should enhance candidate consistency and reduce interviewer-candidate interactions. Effects on validity are mixed. It may decrease contamination, but increase deficiency if it precludes relevant information from emerging. Finally, it could elicit negative user reactions because it restricts freedom and may lead to an awkward conversation. It prevents interviewers from using candidate questions to judge candidates, and it prevents candidates asking questions and using the information to shape their answers (Beatty, 1986).

Research and practice issues. Most importantly, what is the net effect on validity? Is the increase in reliability and decrease in contamination more important than the increase in deficiency? Also, are user reactions sufficiently negative such that this component 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 can be made known in advance. Another possibility is to provide a separate meeting for questions. For example, organizations often conduct socials with college candidates to provide information. Still another possibility is to specifically ask candidates for their questions, and then score question quality as a reflection of candidate preparation.

8. Rate Each Answer or Use Multiple Scales

Explanation and alternatives. There are two elements to this component. First, ratings can be made on each answer or on the entire interview. Second, multiple ratings or only a single rating can be made. Rating each answer is more structured because judgments are more linked to specific responses. Multiple ratings are more structured because they are more extensive. These elements are considered together because they define three common levels of structure, similar to those suggested by Huffcutt and Arthur (1994).

The first and highest level is to rate each answer, typically during the interview with scales tailored to each question. This is used by the highly structured (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, 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 usually 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 at the end. This level is typical of, but not unique to less structured and older approaches (Campbell, Prien, & Brailey, 1960; Ghiselli, 1966; Harris, 1972; McMurry, 1947; Meyer, 1956; Mischel, 1965; Plag, 1961; Pulos, Nichols, Lewinsohn, & Koldjeski, 1962). Many interviews that make dimensional ratings will also make an overall rating or rank the candidates (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 ratings are based on responses to the same questions (Table 1). With ratings of the entire interview, ratings of different candidates (or different interviewers) may be based on different questions. Rating each answer is less cognitively complex because ratings are based on single behavioral events. Ratings during the interview may be less cognitively complex than ratings at the end due to lower memory requirements. There is evidence that decomposed judgments are more reliable than holistic judgments (Armstrong, Denniston, & Gordon, 1975). Finally, higher levels of this

component means making more ratings, thus internal consistency may increase. The meta-analysis by Conway et al. (1995) strongly supports the reliability benefits of multiple ratings.

This component should increase validity. Deficiency may be reduced because more behaviors are evaluated. With specific scales, contamination may be 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 can be overcome. For example, dimension scores are often desired to 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 questions that bear on each dimension. If dimension ratings are preferred, questions on each dimension can be clustered together or indicated (Schmitt & Ostroff, 1986) to reduce complexity. Rather than wait until the end, ratings can be made during the interview as questions on each dimension are completed.

A customized rating scale for each question is not absolutely essential; a common scale can be used for all questions. When different questions are asked across interviews, the applicable questions can be rated and then averaged.

9. Use Detailed Anchored Rating Scales

Explanation and alternatives. This component of structure emerged early (Adams & Smeltzer, 1936; Freeman et al., 1942; Smeltzer & Adams, 1936). Interview rating scales reflect measurement developments in other areas. For example, after research on anchored scales for performance appraisal in the 1970s, nearly all subsequent published interviews had anchored scales.

Anchored rating scales use behavioral examples to illustrate scale points in order to reduce ambiguity and semantic differences possible with adjective anchors (Smith & Kendall, 1963). Modeled after Thurstone (1927), developing such scales involves collecting example answers, judging the goodness of the answers, and selecting unambiguous answers to illustrate points along the scale. Simpler procedures involve intuitively developing example answers.

TABLE 3

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?

Scaled examples	Descriptions	Evaluations	Comparisons
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.			* Next 20% of candidates
3. * 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	* Middle 20% of candidates
2. * I did important tasks first.			*Next 20% of candidates
1. * I did first come first serve, or I asked the supervior.	* 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.

At least four types of anchors have been used (Table 3). 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 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 as well. 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). There are also 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 1). If objectivity increases accuracy, they may also reduce contamination and deficiency. Because anchors are usually in terms of job behaviors or requirements, they also might increase job relatedness.

Two interviewing studies found anchored scales had higher interrater reliability than unanchored scales (Maas, 1965; Vance et al., 1978), and one study also found higher accuracy (Vance et al., 1978). However, extensive research on performance appraisal does not unambiguously support the value of anchored 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. (1996) 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, rather than strong evidence, may be contributing to their popularity. Thus,

future research must clearly determine their effects on reliability and validity.

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 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 can be obtained from actual candidates in pilot interviews, or records of answers to similar questions from previous candidates can be used (Green et al., 1993). Interviewers can be asked about answers from previous candidates (Latham et al., 1980). Brainstorming can be conducted with 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 may enhance structure because it reduces memory decay (M. Campion et al., 1988) and avoids recency and primacy effects (Schmitt & Ostroff, 1986). These benefits may be most apparent when ratings are made at the end or based on multiple questions. Notetaking requires justifying the ratings. This encourages interviewers to attend to answers and to organize their thoughts, thus possibly increasing accuracy.

There are several distinctions. First, notetaking can be extensive or brief. More extensive is more structured, such as summarizing each answer. Second, notetaking can be required or optional, with required being more structured. Third, notetaking can record answers or facts, or it can record evaluations or judgments. The former is more structured because it helps ensure information is perceived accurately. Fourth, notetaking can occur during or after the interview, with during being more structured due to less memory loss.

The highest level of structure is 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 is optional notes or brief notes of answers or evaluations, often at the end (Drucker, 1957; Heneman et

al., 1975; Shaw, 1952; Tarico et al., 1986; Yonge, 1956). The lowest level is no notetaking.

Relatedly, forms can be developed to record notes. 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 cannot be assessed because they are confounded with other components.

Notetaking probably occurred in other interviews but was not considered important enough to mention in the article. Yet, notetaking is very important to interviewers. They 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. Notetaking should make evaluations more consistent, thus increasing test-retest and interrater reliability (Table 1). There should be less disagreement as well.

There is evidence that notetaking enhances recall in interviews (Macan & Dipboye, 1994; Schuh, 1980). It is analogous to using diaries in performance appraisal to help organize information and increase accuracy (DeNisi, Robbins, & Cafferty, 1989). There is evidence that notetaking increases learning (Carrier & Titus, 1979; Kiewra, DuBois, Christian, McShane, Meyerhoffer, & Roskelley, 1991) and helps jurors make correct distinctions (ForsterLee, Horowitz, & Bourgeois, 1994). As such, notetaking may increase accuracy and perhaps job-relatedness. It may reduce deficiency because it helps ensure that important information is recorded and considered, and it may reduce contamination because of its verifiable nature.

User reactions are uncertain. Notetaking may reduce EEO bias by focusing attention on candidate answers and away from illegal factors. Being able to reconstruct answers provides documentation needed for defensibility (Pursell et al., 1980). However, interviewers may find notetaking burdensome. It can be distracting in clinical interviews (Hickling, Hickling, Sison, & Radetsky, 1984). Candidate reactions are ambiguous. Notetaking may reduce eye contact, increase evaluation apprehension, and decrease conversational naturalness (Dipboye & Gaugler, 1993). People prefer counselors who refrain from notetaking (Miller, 1992). Nevertheless, it shows that the interviewer is paying attention to the answers, and they are important enough to record.

Research and practice issues. There is inadequate evidence on reliability and validity. Research should also examine the processes through which it works, such as organizing information, reducing memory decay, or abating rating errors. User reactions are uncertain, as well as the issue of how 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 conscientious administration. Having raters evaluate tapes may enhance structure by minimizing effects of interactions between candidates and interviewers. This might also be cost effective for screening interviews if it reduces travel.

11. Use Multiple Interviewers

Explanation and alternatives. Multiple interviewers may be beneficial for several reasons. Sharing perceptions may help interviewers become aware of irrelevant inferences that are not job related (Arvey & J. Campion, 1982). Multiple interviewers may reduce the impact of idiosyncratic biases among interviewers (M. Campion et al., 1988; Hakel, 1982), and aggregating multiple judgments cancels 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 from different perspectives may increase accuracy (Dipboye, 1992). Finally, using 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). Relative superiority is ambiguous. Panels may be more reliable because interviewers all hear the same answers, but serial interviews may be more valid because they obtain a broader sampling of answers (because different questions are asked to avoid repetition). Second, the number of interviewers influences structure. The upper range may be five (DuBois & Watson, 1950) and nine (Hakel, 1971), with two or three most common.

The higher level of structure 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), or the serial interview (Bobbitt & Newman, 1944; Borman, 1982; Dougherty et al., 1986; Gardner & Williams, 1973; Handyside & Duncan, 1954; Hilton et al., 1955; Huse, 1962; Trankell, 1959). The lower level is one interviewer.

This component is *not* strongly associated with other components of structure. Interviews that are unstructured on other components may still use multiple interviewers. A disproportionately large number of studies on panels are in the public sector. Selecting police is most common (DuBois & Watson, 1950; Flynn & Peterson, 1972; Freeman et al., 1942; Landy, 1976; Lowry, 1994; Reynolds, 1979), but they have been used for many 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 heightened needs for fairness perceptions when staffing government jobs.

Effects on reliability, validity, and user reactions. Scores based on multiple raters should be more reliable than those based on single raters. If interviewers are exposed to the same answers (as in a panel), interrater agreement should be higher. Internal consistency should be higher because more judgments make up the total scores. Interviewer-candidate interactions should be diluted with multiple interviewers.

Deficiency should be reduced because relevant information is less likely to be missed 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).

The meta-analysis by Conway et al. (1995) found panel (vs. separate) interviews were correlated .56 with reliability. Wiesner and Cronshaw (1988) found that among unstructured interviews, panels were more valid than individual interviews (.37 vs. .20). But among structured interviews, there was no real difference (.60 vs. .63). McDaniel et al. (1994) found no difference among unstructured interviews (.33 vs. .34), but a slight advantage for individual formats among structured interviews (.38 vs. .46). Finally, Marchese and Muchinsky (1993) found no correlation between number of interviewers and validity.

The equivocal evidence for the validity benefits of multiple interviewers should be interpreted with three caveats. First, many panels were unstructured on other components, thus reducing validity. The meta-analyses controlled for these other components in only a very gross way. Second, the preponderance of public sector settings for panel interviews may have an unknown effect on validity. Third, there can be process losses with groups (e.g., conformity, conflict, loafing), which may reduce the advantages in some applications.

Multiple interviewers may reduce EEO bias if they reduce the effects of idiosyncratic biases. Systems where interviewer decisions are reviewed by others have been related to positive court outcomes (Gollub-Williamson et al., 1996). Also, multiple interviewers allow members of different races or sexes to be represented, thus enhancing perceptions of fairness (Hakel, 1982).

However, candidate reactions might be negative if panels are stressful. 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. Of course, this may be intentional when stress tolerance is a job requirement, such as in police work (Freeman et al., 1942).

Research and practice issues. A central question is whether multiple interviewers have a positive effect on validity. Unlike previous studies, future studies should avoid confounding with other components of structure. Another question is whether multiple interviewers are needed when structure is high on other components. Finally, research is needed on candidate reactions.

There are many practical issues. For example, should questions be asked by the same panel member, or should they rotate? Should all members take notes, or should one run the interview and the others take notes? Who should ask follow-up questions? Might panels be useful with self-managed 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 and evaluate answers differently. With different interviewers, there is no way to distinguish variance due to rating tendencies among interviewers (e.g., leniency) from true score variance among candidates.

Dreher, Ash, and Hancock (1988) argue that aggregating across interviewers might underestimate validity. They note evidence suggesting interviewers have rating tendencies, and they differ in validity. Therefore, validities aggregated across interviewers will be lower than individual validities due to these rating tendencies. Furthermore, nearly all studies combine interviewer data, thus concealing the problem.

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). Differences in decision strategies have been related to differences in judged interviewer effectiveness (Graves & Karren, 1992). More skilled interviewers may elicit more information and make more accurate judgments (Motowidlo et al., 1992), and interviewers higher in conscientiousness may make 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 a fairly structured interview which may have diminished the 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. A compromise with panel interviews is to keep one member the same (Handyside & Duncan, 1954; Landy, 1976). Few articles mention this component of structure, thus the range of typical practice cannot be ascertained.

Effects on reliability, validity, and user reactions. Using the same interviewers should increase test-retest reliability (Table 1). Variance due to interactions with candidates should be reduced due to less variation in interviewers. Fewer interviewers should reduce contamination due to less variance in response tendencies. But fewer interviewers might increase perceptions of EEO bias if it makes the process appear more idiosyncratic.

Research and practice issues. There is not yet a clear understanding of differences in interviewer validities, thus ideographic research is encouraged. Research might continue to emphasize strategies of cue utilization and interviewer behaviors to learn why such differences occur.

If consistent differences exist, then research should focus on solutions. One solution is interviewing training (see below). Another is selecting better interviewers. Research has not been conducted, but it may be possible to hypothesize attributes distinguishing effective interviewers. For example, experience with the job being staffed might enhance knowledge of job requirements. Dipboye (1992) suggests positive 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, M. Campion et al. (1994) found an interrater reliability of .97. This may be practical in situations where the number of candidates or other concerns make it hard 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 especially applies to panels because the interviewers are all present. It is also especially applicable when interviews are spread out in time.

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 to avoid discussing 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. It should enhance test-retest reliability because it reduces the potential for changing standards and other instrumentation effects (Table 1). However, it might reduce interrater reliability and agreement because it prevents differences in evaluations from being identified and corrected. (Of course, increasing agreement by sharing biases will not help validity.)

This component could enhance validity if it reduces contamination. It may enhance perceptions of procedural justice if it prevents consideration of irrelevant information and the emergence of favorites (M. Campion et al., 1988). But restricting interviewer freedom may create negative reactions.

Research and practice issues. All predicted effects need to be tested. The potential negative effects suggest this component cannot be unequivocally recommended without further evidence. Also, there may be interaction effects with other components. This component may not matter if other components are structured, and negative effects on reliability may be prevented by training.

14. Provide Extensive Interviewing Training

Explanation and alternatives. Training is probably the most common way to improve interviews (Dipboye, 1992). However, training is less of a component itself than a way to ensure other components are implemented correctly. Training interviewers is analogous to training test administrators, which is highly recommended (Society for Industrial and Organizational Psychology, 1987), and interviewing is easily taught (Howard, Dailey, & Gulanick, 1979). Interviews require more

highly skilled administrators than most other selection devices, thus training has been discussed since the early literature (Anderson, 1954; Bolanovich, 1944; Handyside & Duncan, 1954; Hovland & Wonderlic, 1939; Raines & Rohrer, 1955; Wonderlic, 1942).

A content analysis of training reported in the literature revealed the following content. Training often begins with a description of the background and purpose of the interview (Anderson, 1954; Bolanovich, 1944; Walters et al., 1993), followed by a discussion of 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 job relatedness is understood (Pursell et al., 1980). Sometimes trainees complete a job analysis survey (Green et al., 1993). Rapport building is also discussed (Motowidlo et al., 1992; Roth & J. Campion, 1992; Robertson et al., 1990).

Training how to select questions and probes is important if there is interviewer discretion on them (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). Avoiding rating errors is often discussed (Carrier et al., 1990; Maurer & Fay, 1988; Walters et al., 1993).

Notetaking is addressed in many programs (Janz, 1982; Mayfield et al., 1980; Orpen, 1985; Pursell et al., 1980; Robertson et al., 1990). EEO laws and requirements are discussed (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 processes. Lecture and discussion are most common. Many programs use behavioral modeling, role-playing and practice interviews, and then feedback and reinforcement from the trainer and classmates (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 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. 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 discussed above and take a day or two.

Instructions to examinees is a big part of standardization in testing. A less researched topic is interviewee training (Dipboye, 1992). This training could be an orientation session or a handout focusing on what to expect during the interview, how to give good answers (e.g., specific examples), and how to prepare. Although training 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.

Effects on reliability, validity, and user reactions. Training could have many benefits. Trained interviewers should be able to elicit content and evaluate consistently, thus improving test-retest and interrater reliability and agreement (Table 1). They should be able to put candidates at ease, thus enhancing consistency and minimizing differences in interactions. The only negative effect is on internal consistency because many programs encourage avoiding the appearance of halo effects across dimensions. The Conway et al. (1995) meta-analysis supports the positive effect on interrater reliability and no effect on internal consistency.

Although many studies show that structured interviews with training have high reliability, little research demonstrates the unique effects of training. Vance et al. (1978) found that brief rating-error training did not improve reliability. Maurer and Fay (1988) combined frame-of-reference training, which emphasizes consistency of evaluation, with rater-error training, but again found no effects. Neither study is a strong

test because their focus on rating errors is only a small part of interviewing training.

Training focusing on job-related questions and objective scoring should increase job-relatedness and decrease deficiency and contamination. The brief training by Vance et al. (1978) on rating errors had no effect on accuracy, but Dougherty et al. (1986) found that more extensive training including practice interviews with feedback did improve validity. Likewise, Pulakos et al. (1995) found that extensive training improved accuracy.

Training usually emphasizes candidates' legal rights, questions that might be discriminatory, and how bias can enter interviews. This should reduce EEO bias, and reviews of court cases have found supportive evidence (Gollub-Williamson et al., 1996). 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 just because training elicits positive reactions, but because training might help them make 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, yet little is spent on evaluation.

Another issue is the 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 few discretionary behaviors and 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. With a large number of interviewers, it may be more cost effective to have a highly structured interview and a short training program, than a less structured interview and extensive training.

A final issue is whether training decays over time. The importance of following up to ensure proper implementation was recognized early (Wonderlic, 1942). Latham and Saari (1984) found that some interviewers did not score answers 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 asked which was best. Interviewers may add their personal touch, and the variance in implementation may increase over time (Dipboye & Gaugler, 1993).

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

15. Use Statistical Rather than Clinical Prediction

Explanation and alternatives. Different interviewers weigh information differently (Graves & Karren, 1992), thus another way to enhance structure is to use statistical procedures rather than interviewer judgments to combine data (Dipboye, 1992). There was a controversy in the 1950s as to whether statistical methods of combining data were better than clinical methods where expert judgment was used. Interviewing was embroiled in this controversy (Holt, 1958; Rafferty & Deemer, 1950; Trankell, 1959). The evidence from many contexts overwhelmingly favors statistical prediction (Meehl, 1954, 1986; Sawyer, 1966).

This distinction also applies to measurement (Sawyer, 1966). Information can be collected in mechanical ways with little judgment and subjectivity or in clinical ways with intuition and perceptions guiding the process. 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.

Statistical prediction is relevant in three situations. The first involves combining ratings from different questions or dimensions 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, such as differential weights for each rating based on judgment (Adams & Smeltzer, 1936; Janz, 1982; Maas, 1965; Orpen, 1985) or relationships with criteria (Dougherty et al., 1986; Walters et al., 1993).

Other formulas use unit weighting where each rating is given equal weight. It 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 (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).

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 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 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) to make final decisions. This does not refer to withholding ancillary information *before* the interview (component number 6), but to combining information *after* the interview. This can reduce validity because it allows subjectivity. Judgments are often made clinically because the data are not comparable. One approach is to convert scores to percentiles or standard scores and then develop weights (probably through judgment) to combine the data. The only requirement to make the prediction statistical is that the formula be consistently applied.

Effects on reliability, validity, and user reactions. Based on the evidence, reliability should improve (Table 1). Internal consistency may also improve because a formula should increase item-total correlations. Statistical approaches should reduce the likelihood that contaminating information will enter the score or that information will be neglected (i.e., deficiency). Consistency in interview decisions relates to favorable court decisions (Gollub-Williamson et al., 1996) and should, thus, reduce EEO bias. User reactions should be unaffected. 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.

Meta-analyses have yielded mixed results. Conway et al. (1995) found mechanical combination yielded higher reliabilities than subjective methods. But in Wiesner and Cronshaw (1988), panel interviews using averaging had a mean validity of .41 (corrected) compared to .64 using consensus. The consensus interviews also had higher reliabilities

(.74 vs. .84). These results were contrary to their hypothesis and past findings. 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 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 most common. Sometimes differential weighting is desired due to differences in importance between job requirements, but unit weighting can still be used by having the number of questions reflect these differences (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 expectations, and few studies have been conducted. Future research could determine how consensus discussions 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 (M. Campion et al., 1988) or drop deviant ratings (Freeman et al., 1942).

Combining the interview with other information is a very important issue for future research. This is an important opportunity to improve the hiring process through statistical prediction. Some evidence supports the value of statistical methods in combining diverse information in assessment centers (Borman, 1982; Wollowick & McNamara, 1969).

Discussion

Structure and Psychometric Properties

This review supports several conclusions. First, interview structure is complex. This review delineated 15 components, and there may be more that were not identified. Second, any interview could 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 no good rationale for using completely unstructured interviews. We suggest that interviews should be structured in all possible ways within any limits imposed by interviewer and candidate reactions. 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.

There is little consensus among experts as to which components are most important (M. Campion & Palmer, 1995). Research is needed that manipulates the components to determine which are most important and how much structure is required. Nevertheless, based on amount of previous research, strongest psychometric arguments, and 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. 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. Others 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, 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 prohibited from discussing candidates may be resented by interviewers. The interview serves recruiting and public relations roles, in addition to the selection role, and potential trade-offs between psychometric improvements and user reactions must be recognized and avoided or minimized.

On the other hand, these components may not necessarily create negative user reactions. There is simply too little empirical research on user reactions at the present time to know for sure. Furthermore, components with the most negative expected impact may be the least critical to psychometric performance. If a very high level of structure results in some negative reactions, a more moderate level might be acceptable.

Future research could examine which characteristics of participants predispose them to react negatively. For example, candidates with greater impression management needs may object more (Fletcher, 1989). Likewise, highly experienced interviewers may object more because it may 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 enrichment (M. Campion, 1988) of the task (Dipboye, 1995), thus degrading motivation. Interviewers may depart from structure due to boredom. Third, specific motivations influence ratings. For example, hiring quotas or tight labor markets may inflate ratings (Carlson et al., 1971; Webster, 1982), and marginal candidates who will become coworkers may receive very low ratings (Eder, 1989). Finally, accountability may increase motivation. It can increase the accuracy of assessments (Rozelle & Baxter, 1981), and it may be enhanced by simple monitoring (Pulakos, 1995).

Other aspects could be structured to improve user reactions and other outcomes, but have been inadequately studied to draw conclusions. For example, the physical environment such as discomfort (e.g., hard chair), lack of privacy (e.g., interview in public area), and lack of presence (e.g., telephone or computer interviews) may create negative reactions and reduce reliability by distracting participants. The psychological environment such as rapport, clear explanations, and stress may influence candidate reactions, consistency, and interactions with interviewers. Finally, nonverbal and paralanguage may influence outcomes, yet interviews have not been structured to avoid, control, or measure such behaviors. It is unknown whether these behaviors detract from or enhance psychometric properties (Motowidlo & Burnett, 1995), but interviews should avoid inadvertent contamination (Washburn & Hakel, 1973).

Theory Relevant to Structured Interview

Another conclusion is that theory has not played an important role in this area. Past research was very applied. This paper relied on psychometric theory to explain the operation of structured interviews. However, other more content- as opposed to measurement-oriented theories may offer insight. For example, cognitive theory (Lord & Maher, 1991) might be used to consider underlying mechanisms. Structure may reduce information processing requirements and overload, thus allowing interviewers to focus on 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.

Attribution theory (Kelley, 1967) may offer insight. 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 different implicit personality theories (Schneider, 1973) for desirable candidates' qualities. Structure can reduce differences in perspective.

Finally, Webster (1982) describes interviewer decision-making models. One model explains the role of conflict and stress, another explains the role of information processing, and still another explains the role of affect and preferences. Structure might define the decision-making task such that the influence of these processes is 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 length considerations and the fact that many studies focused on other issues. Also, interviews are often just one of many selection procedures examined and may not have been the primary emphasis. Further progress in accumulating knowledge would be enhanced if future studies reported fuller information.

These problems are troubling for meta-analyses. Such techniques can correct for statistical limitations, but not for lack of information, confounded components of structure, or insufficient primary studies.

An equally difficult issue is the unknown construct validity of most 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 vs. job performance) and time orientation (i.e., current identification vs. future prediction). 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.

REFERENCES

Adams CR, Smeltzer CH. (1936). The scientific construction of an interview chart. *Personnel, 13,* 14–19.

Albrecht PA, Glaser EM, 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 RC. (1954). The guided interview as an evaluative instrument. *Journal of Educational Research, 48,* 203–209.

Armstrong JS, Denniston WB Jr., Gordon MM. (1975). The use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance, 14,* 257–263.

Arvey RD, (1995, May). Panelist. In Campion MA, Palmer DK (Chairs), *Taking stock of structure in the employment interview.* Panel discussion presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Arvey RD, Campion JE. (1982). The employment interview: A summary and review of recent research. PERSONNEL PSYCHOLOGY, *35,* 281–322.

Arvey RD, Faley RH. (1988). *Fairness in selecting employees* (2nd ed.). Reading, MA: Addison-Wesley.

Arvey RD, Miller HE, Gould R, Burch P. (1987). Interview validity for selecting sales clerks. PERSONNEL PSYCHOLOGY, *40,* 1–12.

Barrett GV, Svetlik B, Prien EP. (1967). Validity of the job-concept interview in an industrial setting. *Journal of Applied Psychology, 51,* 233–235.

Beatty RN. (1986). *The five-minute interview.* New York: Wiley.

Bender WRG, Loveless HE. (1958). Validation studies involving successive classes of trainee stenographers. PERSONNEL PSYCHOLOGY, *11,* 491–508.

Bobbitt JM, Newman SH. (1944). Psychological activities at the United States Coast Guard Academy. *Psychological Bulletin, 41,* 568–579.

Bolanovich DJ. (1944). Selection of female engineering trainees. *Journal of Education Psychology, 35,* 545–553.

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

Campbell JT. (1962). Assessments of higher-level personnel: 1. Background and scope of the research. PERSONNEL PSYCHOLOGY, *15,* 57–62.

Campbell JT, Prien EP, Brailey LG. (1960). Predicting performance evaluations. PERSONNEL PSYCHOLOGY, *13,* 435–440.

Campion JE, Arvey RD. (1989). Unfair discrimination in the interview. In Eder RW, Ferris GR (Eds.), *The employment interview: Theory, research, and practice* (pp. 61–73). Newbury Park, CA: Sage.

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

Campion MA, Campion JE. (1987). Evaluation of an interviewee skills training program in a natural field experiment. PERSONNEL PSYCHOLOGY, *40,* 675–691.

Campion MA, Campion JE, Hudson JP Jr. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79,* 998–1002.

Campion MA, Palmer DK (Chairs). (1995, May). *Taking stock of structure in the employment interview.* Panel discussion presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Campion MA, Pursell ED, Brown BK. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. PERSONNEL PSYCHOLOGY, *41,* 25–42.

Carlson RE, Schwab DP, Heneman HG III. (1970). Agreement among styles of selection interviewing. *Journal of Industrial Psychology, 5,* 8–17.

Carlson RE, Thayer PW, Mayfield EC, Peterson DA. (1971). Improvements in the selection interview. *Personnel Journal, 50,* 268–275, 317.

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

Carrier MR, Dalessio AT, Brown SH. (1990). Correspondence between estimates of content and criterion-related validity values. PERSONNEL PSYCHOLOGY, *43,* 85–100.

Conway JM, Jako RA, Goodman DF. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579.

Cook TD, Campbell DT. (1979). *Quasi experimentations: Designs and analysis for field settings.* Boston: Houghton Mifflin.

Dalessio AT, Silverhart TA. (1994). Combining biodata test and interview information: Predicting decisions and performance criteria. PERSONNEL PSYCHOLOGY, *47,* 303–315.

Delery JE, Wright PM, McArthur K, Anderson DC. (1994). Cognitive ability tests and the situational interview: A test of incremental validity. *International Journal of Selection and Assessment, 2,* 53–58.

DeNisi AS, Robbins T, Cafferty TP. (1989). Organization of information used for performance appraisals: Role of diary-keeping. *Journal of Applied Psychology, 74,* 124–129.

Dillman DA. (1978). *Mail and telephone surveys: The total design method.* New York: Wiley-Interscience.

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

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

Dipboye RL. (1995, May). Panelist. In Campion MA, Palmer DK (Chairs), *Taking stock of structure in the employment interview.* Panel discussion presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Dipboye RL, Fontenelle GA, Garner K. (1984). Effects of previewing the application on interview process and outcomes. *Journal of Applied Psychology, 69,* 118–128.

Dipboye RL, Gaugler BB. (1993). Cognitive and behavioral processes in the selection interview. In Schmitt N, Borman WC, Associates (Eds.), *Personal selection in organizations* (pp. 135-170). San Francisco: Jossey-Bass.

Dipboye RL, Gaugler BB, Hayes T. (1990, April). *Individual differences among interviewers in the incremental validity of their judgments.* Paper presented at the Fifth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Miami.

Dougherty TW, Ebert RJ, Callender JC. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9–15.

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

Drasgow F, Hulin CL. (1990). Item response theory. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology: Vol. 1* (2nd ed., pp. 577–636). Palo Alto, CA: Consulting Psychologist Press.

Dreher GF, Ash RA, Hancock P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. PERSONNEL PSYCHOLOGY, *41,* 315–327.

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

DuBois PH, Watson RI. (1950). The selection of patrolmen. *Journal of Applied Psychology, 34,* 90–95.

Eder RW. (1989). Contextual effects on interview decisions. In Eder RW, Ferris GR (Eds.), *The employment interview: Theory, research, and practice* (pp. 113–126). Newbury Park, CA: Sage.

Edwards JC, Johnson EK, Molidor JB. (1990). The interview in the admission process. *Academic Medicine, 65,* 167–177.

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.

Feild HS, Gatewood RD. (1989). Development of a selection interview: A job content strategy. In Eder RW, Ferris GR (Eds.), *The employment interview: Theory, research, and practice* (pp. 145–157). Newbury Park, CA: Sage.

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

Fletcher C. (1989). Impression management in the selection interview. In Giacalone RA, Rosenfeld P (Eds.), *Impression management in the organization* (pp. 269–281). Hillsdale, NJ: Earlbaum.

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

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

Freeman GL, Manson GE, Katzoff ET, Pathman JH. (1942). The stress interview. *Journal of Abnormal and Social Psychology, 37,* 427–447.

Gardner KE, Williams APO. (1973). A twenty-five year follow-up of an extended interview selection procedure in the Royal Navy. *Occupational Psychology, 47,* 1–13.

Gehrlein TM, Dipboye RL, Shahani C. (1993). Nontraditional validity calculations and differential interviewer experience: Implications for selection interviews. *Educational and Psychological Measurement, 53,* 457–469.

Ghiselli EE. (1966). The validity of a personnel interview. PERSONNEL PSYCHOLOGY, *19,* 389–394.

Glaser R, Schwarz PA, Flanagan JC. (1958). The contribution of interview and situational performance procedures to the selection of supervisory personnel. *Journal of Applied Psychology, 42,* 69–73.

Gollub-Williamson LR, Campion JE, Malos SB, Roehling MV, Campion MA. (1996). *The employment interview on trial: Linking legal defensibility with interview structure.* Manuscript submitted for publication.

Graves LM, Karren RJ. (1992). Interviewer decision processes and effectiveness: An experimental policy-capturing investigation. PERSONNEL PSYCHOLOGY, *45,* 313–340.

Green PC. (1995, May). Panelist. In Campion MA, Palmer DK (Chairs), *Taking stock of structure in the employment interview.* Panel discussion presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Green PC, Alter P, Carr AF. (1993). Development of standard anchors for scoring generic past-behaviour questions in structured interviews. *International Journal of Selection and Assessment, 1,* 203–212.

Grove DA. (1981). A behavioral consistency approach to decision making in employment selection. PERSONNEL PSYCHOLOGY, *34,* 55–64.

Hakel MD. (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 MD. (1982). Employment interviewing. In Rowland KM, Ferris GR (Eds.), *Personnel management* (pp. 129–155). Boston: Allyn and Bacon.

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

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

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

Harvey RJ. (1991). Job analysis. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology: Vol. 2* (2nd ed., pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.

Heneman HG III, Schwab DP, Huett DL, Ford JJ. (1975). Interviewer validity as a function of interview structure, biographical data, and interviewee order. *Journal of Applied Psychology, 60,* 748–753.

Hickling LP, Hickling EJ, Sison GFP Jr, Radetsky S. (1984). The effect of note-taking on a simulated clinical interview. *Journal of Psychology, 116,* 235–240.

Hilton AC, Bolin SF, Parker JW Jr, Taylor EK, Walker WB. (1955). The validity of personnel assessments by professional psychologists. *Journal of Applied Psychology, 39,* 287–293.

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

Hovland CI, Wonderlic EF. (1939). Prediction of industrial success from a standardized interview. *Journal of Applied Psychology, 23,* 537–546.

Howard GS, Dailey PR, Gulanick NA. (1979). The feasibility of informed pretests in attenuating response-shift bias. *Applied Psychological Measurement, 3,* 481–494.

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

Huffcutt AI, Roth PL, McDaniel MA. (1995, May). *Assessment of mental ability in the employment interview.* Paper presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Huse EF. (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.

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

Kelly EL, Fiske DW. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist, 5,* 395–406.

Kesselman GA, Lopez FE. (1979). The impact of job analysis on employment test validation for minority and nonminority accounting personnel. PERSONNEL PSYCHOLOGY, *32,* 91–108.

Kiewra KA, DuBois NF, Christian D, McShane A, Meyerhoffer M, Roskelley D. (1991). Note-taking functions and techniques. *Journal of Educational Psychology, 83,* 240–245.

Kinicki AJ, Lockwood CA, Hom PW, Griffeth RW. (1990). Interviewer predictions of applicant qualifications and interviewer validity: Aggregate and individual analyses. *Journal of Applied Psychology, 75,* 477–486.

Kleiman LS, Faley RH. (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 ST, Rosa RM. (1984). The applicant interview as a predictor of resident performance. *Journal of Medical Education, 59,* 425–426.

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

Landy FJ, Farr JL. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Langdale JA, 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 GP, Finnegan BJ. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In Schuler H, Farr JL, Smith M (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 41–55). Hillsdale, NJ: Lawrence Erlbaum.

Latham GP, Saari LM. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69,* 569–573.

Latham GP, Saari LM, Pursell ED, Campion MA. (1980). The situational interview. *Journal of Applied Psychology, 65,* 422–427.

Latham GP, Skarlicki DP. (1995). Criterion-related validity of the situational and patterned behavior description interviews with organizational citizenship behavior. *Human Performance, 8,* 67–80.

Locke EA, Latham GP. (1984). *Goal-setting: A motivational technique that works.* Englewood Cliffs, NJ: Prentice-Hall.

Lopez FM Jr. (1966). Current problems in test performance of job applicants: I. PERSONNEL PSYCHOLOGY, *19,* 10–18.

Lord RG, Maher KJ. (1991). Cognitive theory in industrial and organizational psychology. In Dunnette MD, Hough L (Eds.), *Handbook of industrial and organizational psychology: Vol. 2* (2nd ed., pp. 1–62). Palo Alto, CA: Consulting Psychologists Press.

Lowry PE. (1994). The structured interview: An alternative to the assessment center? *Public Personnel Management, 23,* 201–215.

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

Macan TH, Dipboye RL. (1994). The effects of the application on processing of information from the employment interview. *Journal of Applied Social Psychology, 24,* 1291–1314.

Marchese MC, Muchinsky PM. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment, 1,* 18–26.

Maurer SD, 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 EC. (1964). The selection interview: A reevaluation of published research. PERSONNEL PSYCHOLOGY, *17,* 239–260.

Mayfield EC, Brown SH, Hamstra BW. (1980). Selection interviewing in the life insurance industry: An update of research and practice. PERSONNEL PSYCHOLOGY, *33,* 725–739.

McDaniel MA, Whetzel DL, Schmidt FL, Maurer SD. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

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

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

Meehl PE. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370–375.

Meyer HH. (1956). An evaluation of a supervisory selection program. PERSONNEL PSYCHOLOGY, *9,* 499–513.

Miller MJ. (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 JW. (1946). Some notes on oral examinations. *Public Personnel Review, 7,* 15–18.

Motowidlo SJ, Burnett JR. (1995). Aural and visual sources of validity in structured employment interviews. *Organizational Behavior and Human Decision Processes, 61,* 239–249.

Motowidlo SJ, Carter GW, Dunnette MD, Tippins N, Werner S, Burnett JR, Vaughan MJ. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77,* 571–587.

Mumford MD, Stokes GS. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology: Vol. 3* (2nd ed., pp. 61–138). Palo Alto, CA: Consulting Psychologists Press.

Murphy KR, Cleveland JN. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Nevo B, Berman JA. (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 JL. (1944). Improvement of employment interviewing. *Journal of Consultirtg Psychology, 8,* 64–69.

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

Pulakos ED. (1995, May). Panelist. In Campion MA, Palmer DK (Chairs), *Taking stock of structure in the employment interview.* Panel discussion presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Pulakos ED, Nee MT, Kolmstetter EB. (1995, May). Effects of training and individual differences on interviewer rating accuracy. In Kolmstetter EB (Chair), *Interviewer and contextual factors that make a difference in interview validity.* Symposium presented at the Tenth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Orlando.

Pulakos ED, Schmitt N. (1995). Experience-based and situational interview questions: Studies of validity. PERSONNEL PSYCHOLOGY, *48,* 289–308.

Pulakos ED, 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,* 85–102.

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

Pursell ED, Campion MA, Gaylord SR. (1980). Structured interviewing: Avoiding selection problems. *Personnel Journal, 59,* 907–912.

Putney RW. (1947). Validity of the placement interview. *Personnel Journal, 26,* 144–145.

Rafferty JA, Deemer WL Jr. (1950). Factor analysis of psychiatric impressions. *Journal of Educational Psychology, 41,* 173–183.

Raines GN, Rohrer JH. (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 AH. (1979). The reliability of a scored oral interview for police officers. *Public Personnel Management, 8,* 324–328.

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

Roth PL, Campion JE. (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 RM, Baxter JC. (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 PR, Dreher GF. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401–410.

Sackett PR, Wilson MA. (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 DJ. (1973). Implicit personality theory: A review. *Psychological Bulletin, 79,* 294–309.

Schriesheim CA, Solomon E, Kopelman RE. (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 AJ. (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 Fallon BJ, Pfister HP, Brebner J (Eds.), *Advances in industrial organizational psychology* (pp. 343–354). Amsterdam: North-Holland.

Schwab DP, Heneman HG III. (1969). Relationship between interview structure and interviewer reliability in an employment situation. *Journal of Applied Psychology, 53,* 214–217.

Scott WD. (1915, October). The scientific selection of salesmen. *Advertising and Selling,* 5–6, 94–96.

Shahani C, Dipboye RL, Gehrlein TM. (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 CH, Adams CR. (1936). A comparison of graphic and narrative interview reports. *Personnel, 13,* 41–45.

Smith PC, Kendall LM. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Smither JW, Reilly RR, Millsap RE, Pearlman K, Stoffey R. (1993). Applicant reactions to selection procedures. 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.

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

Stohr-Gillmore MK, Stohr-Gillmore MW, 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 VS, Altmaier EM, Smith WL, Franken EA, Berbaum KS. (1986). Development and validation of an accomplishment interview for radiology residents. *Journal of Medical Education, 61,* 845–847.

Thornton GC III, Byham WC. (1982). *Assessment centers and managerial performance.* New York: Academic Press.

Thurstone LL. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

Tinsley HEA, Weiss DJ. (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.

Tubiana JH, 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.

Tucker DH, Rowe PM. (1977). Consulting the application form prior to the interview: An essential step in the selection process. *Journal of Applied Psychology, 62,* 283–287.

Tullar WL. (1989). Relational control in the employment interview. *Journal of Applied Psychology, 74,* 971–977.

Tullar WL, Mullins TW, Caldwell SA. (1979). Effects of interview length and applicant quality on interview decision time. *Journal of Applied Psychology, 64,* 669–674.

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

Vance RJ, Kuhnert KW, Farr JL. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. *Organizational Behavior and Human Performance, 22,* 279–294.

Vernon PE. (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 LA. (1974). The validity of an employment interview independent of psychometric variables. *Australian Psychologist, 9,* 68–77.

Walters LC, Miller MR, Ree MJ. (1993). Structured interviews for pilot selection: No incremental validity. *International Journal of Aviation Psychology, 3,* 25–38.

Wanous JP. (1980). *Organizational entry.* Reading, MA: Addison-Wesley.

Washburn PV, Hakel MD. (1973). Visual cues and verbal content as influences on impressions formed after simulated employment interviews. *Journal of Applied Psychology, 58,* 137–141.

Webster EC. (1959). Decision making in the employment interview. *Personnel Administration, 22* (3), 15–22.

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

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

Wernimont PF, Campbell JP. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376.

Wiener Y, Schneiderman ML. (1974). Use of job information as a criterion in employment decisions of interviewers. *Journal of Applied Psychology, 59,* 699–704.

Wiesner WH, Cronshaw SF. (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 NAB. (1948). The work of the civil service selection board. *Occupational Psychology, 22,* 204–212.

Wollowick HB, McNamara WJ. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53,* 348–352.

Wonderlic EF. (1942). Improving interview techniques. *Personnel, 18,* 232–238.

Wright OR Jr. (1969). Summary of research on the selection interview since 1964. PERSONNEL PSYCHOLOGY, *22,* 391–413.

Wright PM, Lichtenfels PA, Pursell ED. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology, 62,* 191–199.

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

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