Employment Interview on Trial: Linking Interview Structure With Litigation Outcomes

Laura Gollub Williamson and James E. Campion
University of Houston

Mark V. Roehling
Cornell University

Stanley B. Malos
San Jose State University

Michael A. Campion
Purdue University

The authors linked interview structure and litigation outcomes conceptually and empirically. Using legal and psychological literatures, they established a conceptual link based on reduced opportunities for differential treatment through standardization, reduced potential for bias through increased objectivity, and increased job relatedness. Analyzing decisions regarding 84 disparate-treatment claims and 46 disparate-impact claims in federal court cases, they established an empirical link between interview structure and how judges explained their verdicts. The 17 aspects of interview structure were scored in these cases. They collapsed into 3 composites: objective-job related, standardized administration, and multiple interviewers. Most items and composites were significantly related to favorable verdicts for defendants in both types of claims. The objective-job related composite was most highly related, followed by standardized administration. It is concluded that structure enhances interview reliability and validity, and it is also linked to litigation outcomes.

The employment interview is a widely used selection device. Although historically the interview was criticized as having low validity (e.g., Mayfield, 1964; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949; Wright, 1969), more recent reviews and meta-analyses have been more favorable (e.g., Harris, 1989; Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988). They suggest that the interview can be quite valid, especially if (and often only if) it is structured. On the basis of over 100 studies, the meta-analyses have reported validities for structured interviews ranging from .24 to .34 (.44 to .62 corrected), compared with validities for unstructured interviews ranging from .11 to .18 (.20 to .33 corrected).

In contrast, very little research exists on the legal defensibility of the interview as a selection device. Interviews are subject to the same legal standards as other selection devices (Uniform Guidelines; Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978), and the traditional unstructured interview may be highly vulnerable to legal challenge because of potential bias made possible by its subjective and inconsistent nature (Arvey, 1979; Arvey & Campion, 1982; Arvey & Faley, 1988; J. E. Campion & Arvey, 1989). On the other hand, structured interviews may be more resistant to legal challenge because of their heightened objectivity, standardization, and job relatedness (M. A. Campion, Pursell, & Brown, 1988; Pursell, Campion, & Gaylord, 1980).

The purpose of our study was to investigate the latter thesis by (a) examining the conceptual link between interview structure and litigation outcomes in discrimination cases and (b) evaluating empirically the relationship between interview structure and federal trial court judges' decisions in reported employment discrimination cases.

Conceptual Link Between Interview Structure and Litigation Outcomes

Subjective Hiring Procedures

Although subjective hiring procedures, such as traditional unstructured selection interviews, are not prohibited per se, the use of subjective procedures is subject to
close scrutiny by the courts (Larson, 1995; see also Lee v. Russell County Board of Education, 1982). Generally, the more subjective a defendant's proffered reasons for a challenged employment decision, the more difficult will be the employer's task of meeting its burden of producing rebuttal evidence (Robbins v. White-Wilson Medical Clinic, Inc., 1981).

Two distinct concerns regarding the use of subjective hiring procedures have been articulated by the courts. First, courts have recognized that subjective processes are susceptible to inappropriate biases. Subjective hiring practices have been described as a "ready mechanism" for racial discrimination (Sims v. Montgomery County Commissioners, 1982, p. 426; see also Phillips v. Amoco Oil Company, 1982; Harris v. Birmingham Board of Education, 1983).

Second, courts may unfairly deny plaintiffs a reasonable opportunity to challenge and rebut the claimed basis for the employer's decision (Lee v. Russell County Board of Education, 1982; cf. Conner v. Gordon Bus Company, 1985). In other words, totally subjective procedures are sufficiently ambiguous that they do not allow plaintiffs an adequate way to scrutinize the basis of the employment decision. This may be characterized as denying the plaintiff due process.

**Interview Structure**

Some previous research on structured interviewing was explicitly motivated by a desire to enhance the legal defensibility of hiring decisions (e.g., M. A. Campion et al., 1988; Pursell et al., 1980). Our definition of structure used here is based on that of M. A. Campion, Palmer, and J. E. Campion (1997). It considers structure 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. Enhancements may improve either the content of the interview or the evaluation process. Psychometric properties refer to both reliability and validity, but enhancements that create positive user reactions are also included. We used this broad definition of structure because of the wide range of enhancements to the interview observed in the existing literature (M. A. Campion et al., 1997) and because this will help avoid overlooking any enhancements that may be relevant to legal defensibility. Illustrations follow.

Some enhancements are intended to improve the objectivity or job relatedness of the interview. For example, basing interview questions on a job analysis can help ensure job relatedness. Job analysis is a key recommendation of the Uniform Guidelines, and it can help support the content validity of the interview (Arvey & Faley, 1988). Job analysis may prevent interviewers from basing questions on idiosyncratic beliefs about job requirements (Dipboye, 1994), and it has been shown to reduce the likelihood of bias (Kesselman & Lopez, 1979). It has also been associated with favorable court decisions on other selection procedures (Kleiman & Faley, 1985), and it may do the same for interviews (J. E. Campion & Arvey, 1989).

More objective, specific, and behaviorally oriented questions and evaluation criteria can reduce bias by minimizing the ambiguity, interpretation required, and overall subjectivity of interview decisions. Evidence of test fairness for interview questions with these characteristics has been found (M. A. Campion et al., 1988; Pulakos & Schmitt, 1995).

Interviewer training may also improve objectivity and job relatedness (e.g., Arvey, 1979; J. E. Campion & Arvey, 1989; Dipboye, 1992; Schmitt, 1976). In order to reduce bias, a key part of most training programs is a review of equal employment opportunity laws, and interviewers are typically trained to avoid irrelevant or potentially discriminatory questions. They are also trained to base rating decisions on job-related information, to avoid stereotypes, and so on.

Other enhancements are intended to improve the interview by standardizing the administration process. For example, using the same questions with all candidates can reduce bias because of the obvious fairness of giving all the candidates the exact same opportunity. This type of structure limits many sources of potential bias in the interview (Dipboye, 1994; Dipboye & Gaugler, 1993) and has been viewed as defensible by attorneys (Latham & Finnegan, 1987). Likewise, rating scales can reduce bias and increase reliability by making judgments more specific and behavioral. Further, their quantitative nature allows more precise comparisons and reference points that help interviewers be more consistent across candidates (Arvey & Faley, 1988).

Finally, some enhancements are intended to improve the interview by using multiple interviews. Multiple interviewers or panels can reduce bias because ratings are based on more observations. The effects of idiosyncratic opinions may be canceled out (Arvey & Campion, 1982), and the range of perspectives might increase accuracy (Dipboye, 1992). In addition, using multiple interviewers allows decisions to be reviewed by others before being implemented.

It should be recognized that although the procedural justice literature and the structured interviewing literature have developed independently (partly because the application of justice models to selection has been fairly recent), they both encourage many of the same types of structure. For example, some of these models encourage job relatedness, opportunity to perform, consistency of administration, and propriety of questions (e.g., Gilliland, 1993).
Linking Interview Structure With Legal Defensibility: Standardization, Objectivity, and Job Relatedness

The essence of any disparate treatment claim is that the aggrieved party was treated differently on the basis of a protected characteristic (e.g., race, sex, etc.). To the extent that structure standardizes interview practices, thus treating all job candidates the same, it should both reduce the likelihood that disparate treatment will occur and increase the likelihood that the employer will be able to successfully defend itself if disparate treatment is claimed. Indeed, if all candidates are in fact treated identically, disparate treatment cannot occur—almost by definition.

The increased objectivity associated with structured interviews suggests that they may not be scrutinized by the courts as closely as traditional interviews that involve much greater interviewer subjectivity. Accordingly, it should be easier for an employer to discharge its burden of articulating a nondiscriminatory reason for the hiring decision and producing some evidence in support of the reason. The increased job relatedness of structured interviews is not essential to defending disparate treatment claims, however. An employer does not necessarily need to demonstrate job relatedness in order to discharge its burden of articulating a legitimate, nondiscriminatory reason for the decision. Of course, job relatedness is one very good type of legitimate, nondiscriminatory reason.

In contrast, it is the increased job relatedness of structured interviews that is critical to their greater expected defensibility in disparate impact claims. Once the plaintiff establishes a prima facie case of disparate impact, the defendant must persuade the trier of fact that the challenged practice is job related; merely articulating a legitimate, nondiscriminatory reason is not sufficient.

The defendants should identify those characteristics associated with successful applicants. Guidelines should be developed, for use by interviewers. While it may be impossible and undesirable to eliminate all subjective aspects of the interview, the use of enunciated standards, tailored to the demands of the entry level program and uniformly applied by all interviewers, will guard against abuses of judgment.

Empirical Link Between Interview Structure and Litigation Outcomes

The foregoing analysis strongly suggests that there should be a positive relationship between elements of interview structure and judges' decisions. However, it does not describe the exact components of interview structure that have been discussed in court cases nor their relative importance. This consideration, and the benefits of assessing phenomena using multiple methods, encouraged an empirical assessment of the link between interview structure and legal outcomes. Therefore, the written opinions of federal court judges who decided cases involving the employment interview were examined. This combination of legal and empirical analysis should complement previous narrative reviews of court cases on the interview (cf. Arvey & Faley, 1988; J. E. Campion & Arvey, 1989).

Roehling's (1993) critique of policy-capturing research on judicial decision making raised a number of specific concerns that have been addressed in the present study.

To qualitatively illustrate how judges evaluate interviews, quotations are presented from two cases. In Ste. Marie v. Eastern Railroad Association (1978), the employer was found guilty of discriminating against women by not hiring them for, or promoting them to, managerial or professional positions. The judge noted the absence of numerous aspects of interview structure in explaining the verdict:

The job descriptions are ad hoc and home-made... Not every candidate was asked the same questions; and the answers of the winning candidate were not scored... There are no written criteria... These are... highly subjective and imprecise criteria subject to greatly fluctuating evaluations based upon the background and personality of the evaluator and the chemistry which developed in the interaction with the applicant... A selection process such as that currently employed by defendant fails to conform to Title VII requirements because it is totally discretionary and not guided by written instructions... The lack of objective criteria, the total discretion given the department head to make the choice, and the absence of any restraint to prevent discrimination clearly suffice to establish by a fair preponderance of the evidence that defendant has and is subjecting its female employees to employment discrimination because of their sex. (pp. 676–677, 682, 684)

In Reynolds v. Sheet Metal Workers Local 102 (1980), the court restrained the defendant from using subjective interviews to select applicants to an apprenticeship program. The court found that the interview had many of the subjective procedures that have been "condemned" by the courts, including "broad and undefined criteria... no guidelines... the interviewer's judgments are unviewable... defendants have not attempted to validate any aspect of the interview" (p. 974). The court went on to provide the defendant guidance that, in essence, called for increased structure:

The defendants should identify those characteristics associated with successful applicants... Guidelines should be developed, for use by interviewers... While it may be impossible and undesirable to eliminate all subjective aspects of the interview, the use of enunciated standards, tailored to the demands of the entry level program and uniformly applied by all interviewers, will guard against abuses of judgment (p. 974).

Illustrative Cases

To qualitatively illustrate how judges evaluate interviews, quotations are presented from two cases. In Ste. Marie v. Eastern Railroad Association (1978), the employer was found guilty of discriminating against women by not hiring them for, or promoting them to, managerial or professional positions. The judge noted the absence of numerous aspects of interview structure in explaining the verdict:

The job descriptions are ad hoc and home-made... Not every candidate was asked the same questions; and the answers of the winning candidate were not scored... There are no written criteria... These are... highly subjective and imprecise criteria subject to greatly fluctuating evaluations based upon the background and personality of the evaluator and the chemistry which developed in the interaction with the applicant... A selection process such as that currently employed by defendant fails to conform to Title VII requirements because it is totally discretionary and not guided by written instructions... The lack of objective criteria, the total discretion given the department head to make the choice, and the absence of any restraint to prevent discrimination clearly suffice to establish by a fair preponderance of the evidence that defendant has and is subjecting its female employees to employment discrimination because of their sex. (pp. 676–677, 682, 684)

In Reynolds v. Sheet Metal Workers Local 102 (1980), the court restrained the defendant from using subjective interviews to select applicants to an apprenticeship program. The court found that the interview had many of the subjective procedures that have been "condemned" by the courts, including "broad and undefined criteria... no guidelines... the interviewer's judgments are unviewable... defendants have not attempted to validate any aspect of the interview" (p. 974). The court went on to provide the defendant guidance that, in essence, called for increased structure:

The defendants should identify those characteristics associated with successful applicants... Guidelines should be developed, for use by interviewers... While it may be impossible and undesirable to eliminate all subjective aspects of the interview, the use of enunciated standards, tailored to the demands of the entry level program and uniformly applied by all interviewers, will guard against abuses of judgment (p. 974).

Empirical Link Between Interview Structure and Litigation Outcomes

The foregoing analysis strongly suggests that there should be a positive relationship between elements of interview structure and judges' decisions. However, it does not describe the exact components of interview structure that have been discussed in court cases nor their relative importance. This consideration, and the benefits of assessing phenomena using multiple methods, encouraged an empirical assessment of the link between interview structure and legal outcomes. Therefore, the written opinions of federal court judges who decided cases involving the employment interview were examined. This combination of legal and empirical analysis should complement previous narrative reviews of court cases on the interview (cf. Arvey & Faley, 1988; J. E. Campion & Arvey, 1989).

Roehling's (1993) critique of policy-capturing research on judicial decision making raised a number of specific concerns that have been addressed in the present study.
First, the quantitative analysis of the judicial decisions was supplemented with a traditional legal and conceptual discussion of the issues (as presented above) to support the expected relationships and place results in context. Second, a clear distinction was made between disparate impact and disparate treatment claims, and they are coded and analyzed separately. Third, special care was taken to code only aspects of interview structure and not the judge's conclusions that have explicit legal significance. For example, the existence of a job analysis was coded but not whether the judge concluded that the interview was job related. The latter conclusion has formal legal significance. If the judge concludes that an interview is job related in a disparate impact case, it necessitates a verdict for the defendant (unless the plaintiff can prove there is an alternative practice with less adverse impact).

Fourth, a wide range of factors were controlled, including the law under which the claim was filed and the type of discrimination alleged, as well as the nature of the employer, the job, and the employment decision. In addition, a key informal factor, the judge's political party affiliation, was controlled. Affiliation has been related to liberalism versus conservatism in judicial decision making across a wide spectrum of cases (Aliotta, 1988; Hall & Brace, 1992; Songer & Davis, 1989). Democratic, as opposed to Republican, judges are likely to support more liberal positions in civil rights cases (Gottschall, 1986; Songer & Davis, 1989) and thus decide in favor of plaintiffs. Fifth, the limits of the sample used were identified to minimize the risk that the results will be overgeneralized. Sixth, consideration was given to the appropriateness of aggregating cases over time. As described in a subsequent section, cases were collected from 1972 to 1995. Unlike some areas of law, the fundamental nature of discrimination claims under Title VII did not change during this period. The basic elements of disparate treatment claims were enunciated in McDonnell Douglas Corporation v. Green (1972), and the basic elements of disparate impact claims were established in Griggs v. Duke Power (1971). Although there have been judicial interpretations that purported to affect the administration of the burden of proof in Title VII cases (e.g., Wards Cove Packing Co. v. Atomics, 1989), and a legislative initiative that sought to clarify and codify nuances in the administration of the burden (i.e., Civil Rights Act of 1991), the basic elements of the burden of proof have remained essentially the same.

**Method**

**Selection of Cases**

Judges' written opinions in federal district court cases involving an allegation of discrimination in an employment interview were the data for this study. Case selection had dual goals of maximizing sample size and quality of data. To maximize sample size, several sources were used to locate cases. Preliminarily, the literature discussed above and several legal handbooks on employment discrimination were searched for citations to potentially usable cases. However, the primary method for locating cases was a computer search of federal court cases up to 1995 listed on the LEXIS electronic database. In addition, the comprehensive indexes of the two main reporters of discrimination cases, Employment Practices Decisions (published by Commerce Clearing House) and Fair Employment Practices Cases (published by the Bureau of National Affairs), were manually searched.

To maintain the quality of the data, cases were included in the sample only if they met five criteria. First, use of an employment interview had to be the primary basis for some personnel decision resulting in a charge of discrimination. Second, adequately detailed information about the interview had to be provided in the case. Third, cases had to be tried in the federal district court system. Cases tried at the federal level are assumed to be of broader interest and applicability than those tried in state courts. Although cases are often appealed and district court rulings are sometimes reversed, the decision of the district court is of greatest interest in this study because it bears actual evidence and evaluates that evidence with respect to the complaint. Thus, the district court's focus, like this study's focus, is on the facts of the case, whereas appellate courts tend to examine mostly procedural or legal issues. Fourth, cases had to result in a decision addressing the substantive merits of a discrimination claim. Cases that addressed only procedural issues (e.g., whether a proposed class should be certified in a class action suit) were not included. Fifth, cases had to result in a verdict, for either the plaintiff or the defendant, on a substantive employment discrimination issue. Cases that were decided on procedural grounds were not included, as they were not concerned with whether an interview resulted in unfair discrimination.

**Description of Sample**

A total of 99 federal district court cases involving alleged discrimination in an employment interview were included in this sample (citations in the Appendix). This included 84 disparate-treatment claims and 46 disparate-impact claims. Some cases involved both types of claims, but controlling for these cases did not affect the results. Statistical power was assessed using the $p < .10$ significance level in order to balance Type I and II errors with a moderate sample size. Power levels were 88% and 66% to detect a correlation of 0.30 for disparate-treatment and disparate-impact claims, respectively ($p < .10$, one-tailed; Cohen, 1977).

Judgments were granted for the defendant in 51% of the disparate-treatment claims and in 54% of the disparate-impact claims. Title VII of the Civil Rights Act of 1964 was the basis for 91% of cases. Cases also involved violations of the Civil Rights Acts of 1866 (23%) or 1871 (10%), the Age Discrimination in Employment Act (4%), or some other relevant law (6%). Race discrimination was asserted in 57% and sex discrimination in 34% of cases. Additionally, 9% involved national origin, 4% age, 1% a physical disability, and 2% reverse discrimination. The numbers do not sum to 100% because several cases involved multiple charges.
All judgments were granted between 1972 and 1995, with most occurring between 1977 and 1983 (56%). A single plaintiff was involved in 54% of cases, multiple plaintiffs in 10%, and class actions in 36%. The defendant employer was a private business without a union in 40% of cases, a private business with a union in 7%, public school in 15%, police or fire department in 8%, and some other public organization in 30%. Unskilled jobs were at issue in 45% of cases, skilled jobs in 31%, professional jobs in 21%, and managerial jobs in 23%. New hires were involved in 64% of cases, promotions in 32%, job assignments in 10%, job transfers in 10%, training in 4%, and other employment decisions in 8%. Several cases involved more than one type of job or employment decision.

The judges' political affiliations were 54% Republican and 46% Democratic. This was determined on the basis of the political affiliation of the president appointing the judge, which is a widely used measure (e.g., Carp & Rowland, 1983; Gottschall, 1986; Hall & Brace, 1992; Songer & Davis, 1989). Judges were 96% male and appointed on average in 1981 ($SD = 5.18$).

**Coding Procedure**

The criterion was the district court's verdict. In cases involving decisions on both disparate-treatment and disparate-impact claims, each decision was treated as a separate unit of analysis. The criterion was coded as a dichotomous variable, for plaintiff (indicating discrimination, coded 0) or for defendant (indicating inadequate proof of discrimination, coded 1). Cases were also coded on the descriptive items discussed in the previous section.

The remainder of the coding procedure pertained to the interview itself. Each case was coded using rating scale items describing aspects of interview structure that were developed through the review of the literature and through a preliminary reading of a subset of the cases. Any feature of the interview that appeared to enhance its reliability or validity, either through improved content or evaluation processes, was coded as an aspect of structure (M. A. Campion et al., 1997). Items were developed with an iterative deductive-inductive process whereby items initially based on the literature were tested on a sample of cases, and appropriate modifications were made. Less frequently, items were suggested from reading the cases and then evaluated on the basis of their correspondence to constructs in the literature. This coding system was refined several times and reviewed independently by colleagues before being implemented. The intent was to satisfy three goals simultaneously: (a) to accurately reflect as many of the components of interview structure identified in the literature as possible, (b) to capture all the aspects of structure that were discussed in the cases, and (c) to develop a scoring system that could be reliably administered and that could detect a reasonable level of variance across cases. This process resulted in the development of 17 items.

All items were rated on 5-point rating scales that were anchored with verbal descriptions. Anchors were written to capture discernible levels of meaningful differences between cases on the basis of an iterative process of rating a sample of cases and modifying the anchors. Each scale was ordinal, such that higher values reflected more structured interviews. For example, the following scale was used for the item on specificity of criteria in the interview: 1 = no criteria specified—just overall rating; 2 = general, global, vague criteria; 3 = mixed general and specific criteria; 4 = mostly specific criteria; and 5 = specific criteria with appropriate anchors or examples. Scores were assigned on the basis of the information in the judge's written opinion in the case. In some instances, it was not possible to rate a case on a given item because of the lack of detail in the written opinion. Rather than make poorly supported inferences, these items were not coded and were instead treated as missing.

Such a strategy should work against finding effects, thus providing a conservative test of the hypothesis.

The 17 items were grouped into three composites based on factor analytic results and conceptual similarity of the items. Principal components analysis were performed with varimax rotation. Three factors were retained based on scree plot and interpretability, collectively explaining 65% of the total variance. Items were allocated to factors according to their highest loadings. One item was placed on the factor with its second highest loading (.61 versus .66) in order to make the composites more conceptually homogeneous. Composites scores were computed by averaging the completed items (e.g., if 5 were completed out of 7 items, then the composite would be the average of the 5).

The resulting composites had good conceptual meaningfulness and acceptable reliabilities. The three composites were objective-job related (seven items, coefficient $\alpha = .83$), standardized administration (seven items, $\alpha = .88$), and multiple interviewers (three items, $\alpha = .67$). Abbreviated item descriptions in Tables 1 and 2 summarize the content of each composite (for disparate-treatment and -impact claims, respectively). Sample sizes in the tables indicate that there was a wide range in the number of cases that contained each type of information, but 79% of the items applied to at least half the cases. Total-sample means and standard deviations in Tables 1 and 2 suggest that the full ranges of the scales were used in the coding process. Table 3 shows descriptive statistics for the composites and their intercorrelations. The composites show good range and variation and moderate independence.

Because all of the items in a given composite were missing in only a small number of cases (2%), mean substitution of composite scores was used in those cases (Roth, 1994) so that all analyses could be based on a uniform sample size and the maximum amount of data. Analyses of relationships between the composites and the courts' decisions were conducted with and without various amounts of missing data. Specifically, analyses were conducted without mean substitution for missing composites, and analyses were conducted excluding composites with excessive missing items (e.g., two items missing, half the items missing, etc.). Magnitudes of the relationships increased in all analyses when cases with missing data were excluded, although usually only slightly. Therefore, analyses are reported on the basis of all possible data in order to avoid the loss of statistical power and to use all the relevant data that are available.

**Assessment of Coding Reliability**

The initial sample included cases through 1989 ($N = 89$). In order to assess the reliability of coding, 25 cases were randomly selected and assigned each to one of three independent duplicate...
Table 1
Item-Level Descriptive Statistics for Disparate-Treatment Claims

<table>
<thead>
<tr>
<th>Composite-item</th>
<th>Total sample</th>
<th>Verdict for plaintiff</th>
<th>Verdict for defendant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Objective-job related</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective criteria</td>
<td>81</td>
<td>2.15</td>
<td>0.98</td>
</tr>
<tr>
<td>Behavior- (vs. trait-) based criteria</td>
<td>73</td>
<td>2.77</td>
<td>0.95</td>
</tr>
<tr>
<td>Specific (vs. general) criteria</td>
<td>64</td>
<td>2.39</td>
<td>0.99</td>
</tr>
<tr>
<td>Interviewer experienced or trained</td>
<td>51</td>
<td>3.04</td>
<td>0.60</td>
</tr>
<tr>
<td>Job description or job analysis</td>
<td>53</td>
<td>2.83</td>
<td>1.12</td>
</tr>
<tr>
<td>Interviewer familiar with requirements</td>
<td>64</td>
<td>3.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Validation evidence</td>
<td>10</td>
<td>1.80</td>
<td>1.40</td>
</tr>
<tr>
<td>Standardized administration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of rating form</td>
<td>28</td>
<td>2.38</td>
<td>1.21</td>
</tr>
<tr>
<td>Guides for conducting interview</td>
<td>38</td>
<td>2.24</td>
<td>1.40</td>
</tr>
<tr>
<td>Minimal interviewer discretion</td>
<td>44</td>
<td>2.11</td>
<td>1.43</td>
</tr>
<tr>
<td>Standardized questions</td>
<td>41</td>
<td>2.17</td>
<td>1.39</td>
</tr>
<tr>
<td>Plaintiff's interview not different</td>
<td>45</td>
<td>3.36</td>
<td>1.77</td>
</tr>
<tr>
<td>Interviewer is from same race/sex group as plaintiff</td>
<td>69</td>
<td>1.67</td>
<td>1.44</td>
</tr>
<tr>
<td>Statistical combination of ratings</td>
<td>47</td>
<td>1.68</td>
<td>1.40</td>
</tr>
<tr>
<td>Multiple interviewers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of multiple interviewers</td>
<td>82</td>
<td>3.10</td>
<td>1.00</td>
</tr>
<tr>
<td>Panel interview</td>
<td>65</td>
<td>2.86</td>
<td>1.00</td>
</tr>
<tr>
<td>Interviewer's decision reviewed</td>
<td>67</td>
<td>2.04</td>
<td>1.32</td>
</tr>
</tbody>
</table>

Note. All measures on 5-point scales, with 5 highest.
*p < .10. **p < .05.

coders. Each duplicate coder was provided with training on using the coding form, including example practice cases to code. Correlations between the primary coder and the duplicate coder on the composites ranged from .72 to .90 (all significant at p < .05), indicating an acceptable level of interrater reliability. Examination of scatterplots between pairs of ratings revealed no unusual patterns. Mean differences between the primary coder and the duplicate coder on the composites ranged from

Table 2
Item-Level Descriptive Statistics for Disparate-Impact Claims

<table>
<thead>
<tr>
<th>Composite-item</th>
<th>Total sample</th>
<th>Verdict for plaintiff</th>
<th>Verdict for defendant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Objective-job related</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective criteria</td>
<td>45</td>
<td>2.04</td>
<td>0.95</td>
</tr>
<tr>
<td>Behavior- (vs. trait-) based criteria</td>
<td>40</td>
<td>2.58</td>
<td>0.93</td>
</tr>
<tr>
<td>Specific (vs. general) criteria</td>
<td>35</td>
<td>2.49</td>
<td>0.89</td>
</tr>
<tr>
<td>Interviewer experienced or trained</td>
<td>36</td>
<td>3.06</td>
<td>0.67</td>
</tr>
<tr>
<td>Job description or job analysis</td>
<td>32</td>
<td>2.78</td>
<td>1.18</td>
</tr>
<tr>
<td>Interviewer familiar with requirements</td>
<td>44</td>
<td>3.71</td>
<td>0.76</td>
</tr>
<tr>
<td>Validation evidence</td>
<td>14</td>
<td>1.64</td>
<td>1.43</td>
</tr>
<tr>
<td>Standardized administration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of rating form</td>
<td>16</td>
<td>2.81</td>
<td>1.05</td>
</tr>
<tr>
<td>Guides for conducting interview</td>
<td>30</td>
<td>2.07</td>
<td>1.34</td>
</tr>
<tr>
<td>Minimal interviewer discretion</td>
<td>25</td>
<td>1.88</td>
<td>1.36</td>
</tr>
<tr>
<td>Standardized questions</td>
<td>24</td>
<td>2.25</td>
<td>1.59</td>
</tr>
<tr>
<td>Plaintiff's interview not different</td>
<td>19</td>
<td>3.42</td>
<td>1.84</td>
</tr>
<tr>
<td>Interviewer is from same race/sex group as plaintiff</td>
<td>33</td>
<td>1.42</td>
<td>1.09</td>
</tr>
<tr>
<td>Statistical combination of ratings</td>
<td>28</td>
<td>1.93</td>
<td>1.68</td>
</tr>
<tr>
<td>Multiple interviewers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of multiple interviewers</td>
<td>44</td>
<td>3.14</td>
<td>1.00</td>
</tr>
<tr>
<td>Panel interview</td>
<td>33</td>
<td>2.85</td>
<td>1.00</td>
</tr>
<tr>
<td>Interviewer's decision reviewed</td>
<td>38</td>
<td>2.05</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Note. All measures on 5-point scales, with 5 highest.
*p < .10. **p < .05.
Table 3
Means, Standard Deviations, and Correlations Among the Composites

<table>
<thead>
<tr>
<th>Measure</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Objective-job related</td>
<td>2.83</td>
<td>0.68</td>
<td>--</td>
<td>.71*</td>
<td>.46**</td>
<td>.51**</td>
<td>2.75</td>
<td>0.72</td>
</tr>
<tr>
<td>2. Standardized admin</td>
<td>2.13</td>
<td>1.17</td>
<td>.52**</td>
<td>--</td>
<td>.38**</td>
<td>.28**</td>
<td>2.06</td>
<td>1.15</td>
</tr>
<tr>
<td>3. Multiple interviewers</td>
<td>2.75</td>
<td>0.93</td>
<td>.26**</td>
<td>.27**</td>
<td>--</td>
<td>.18</td>
<td>2.76</td>
<td>0.96</td>
</tr>
<tr>
<td>4. Verdict</td>
<td>0.51</td>
<td>0.50</td>
<td>.38**</td>
<td>.27**</td>
<td>.17*</td>
<td>--</td>
<td>0.54</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Note. Verdict scored 0 for plaintiff and 1 for defendant. All other measures on 5-point scales, with 5 highest. Disparate-treatment claims to left of diagonal (n = 84); disparate-impact claims to right of diagonal (n = 46).

*p < .10. **p < .05.

-0.22 to 0.55 on the 5-point scales (all nonsignificant), indicating an acceptable overall level of interrater agreement.

The initial sample was updated in 1995. A new search of the legal literature was conducted in exactly the same manner as the original search, and 10 subsequent cases were identified that met the inclusion criteria. Because a new coder was used, a random sample of 10 cases from the original sample were recoded. Interrater reliability ranged from a correlation of .48 to .96, and interrater agreement ranged from a mean difference of -.40 to .25 on the 5-point scales. The additional 10 cases were added to the original sample for a total of 99 cases.

Results
Univariate Analyses
Table 1 shows that 11 of the 17 items of interview structure were significantly higher for disparate-treatment claims where the verdict resulted in favor of the defendant. Most of the items in the objective-job related and standardized administration composites were significant. Only 1 of the 3 items in the multiple interviewers composite was significant, however.

Table 2 shows nearly identical results for the disparate-impact claims, with 10 of the 17 items significant. In fact, for 16 of the 17 items, the results are the same (either significant or nonsignificant, and in the same direction) for both disparate-treatment and -impact claims.

Table 3 shows that all three of the composite measures of interview structure were significantly and positively related to a verdict for the defendant in disparate-treatment claims, and all but the multiple interviewers composite were significant for disparate-impact claims. The objective-job related composite was more highly related to the verdict than the other composites in both types of claims, but the differences were not significant.

It might have been expected that the objective-job related composite would be more highly related to the verdict in disparate-impact claims than in disparate-treatment claims because of the need to show job relatedness to defend such claims. This was the trend, but the results were not significant. Likewise, it might have been expected that the standardized administration composite would be more highly related to the verdict in disparate-treatment than in disparate-impact claims because of the need to show consistency to defend such claims. Again, this was not found.

Multivariate Analyses
Regressions were conducted to control for the relationships among the composites. Logistic regression was used because verdict was a dichotomous criterion (Aldrich & Nelson, 1984). As shown in Table 4 (Model 1) for disparate-treatment claims, the $\chi^2$ was significant, indicating that at least one of the predictors was related to the verdict (Demaris, 1992). However, only the coefficient for the objective-job related composite was significant when all three composites were considered together. Both the relative size of the parameter estimate (e.g., standardized $\lambda = 0.41$) and the odds ratio (2.93) for this composite show its relative importance in explaining criterion group membership (Tabachnick & Fidell, 1989).

Note that the odds ratio reflects the chances of a verdict for the defendant divided by the chances of a verdict for the plaintiff. An odds ratio greater than 1.0 means the verdict was more likely for the defendant, whereas an odds ratio less than 1.0 means the verdict was more likely for the plaintiff. Thus, the results indicate that if the interview was highly objective and job related, then it was much more likely that the verdict was in favor of the defendant.

Similar results were obtained for disparate-impact claims, as shown in Table 5. Only the objective-job related composite was significant when all three composites were considered together. The parameter estimate (standardized $\lambda = 0.96$) and the odds ratio (10.96) for this composite appear quite large, indicating the importance of this aspect of structure.

Discriminant analyses were conducted to determine the
degree to which information on interview structure could correctly classify a verdict. The discriminant function for disparate-treatment claims was significant, Wilk's $\lambda = 0.85$, $F(3, 80) = 4.84, p < .05$. Also, 67% of all claims could be correctly classified on the basis of interview structure information in the judge's written opinion. The canonical $R$ was .39. Loadings on the discriminant function and $t$ tests between groups (not shown) indicated that the objective-job related composite was a somewhat stronger explainer of the classification, although the standardized administration and multiple interviewers composites were also important.

Very similar results were obtained for disparate-impact claims, Wilk's $\lambda = 0.72$, $F(3, 42) = 5.41, p < .05$. The percentage of correct classification was 74%, and the canonical $R$ was .53. The objective-job related composite was the most important explainer (not shown), followed by standardized administration. The multiple interviewer composite was not significant.

**Control Variable Analyses**

The case description measures were correlated with the court decisions to determine whether any were significantly related to the verdicts ($p < .10$) and should be considered as control variables. Disparate-treatment

---

**Table 4**

*Logistic Regression Analysis of Verdict on Interview Structure for Disparate-Treatment Claims*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model 1 (Structure only)</th>
<th>Model 2 (Controls added)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Lambda^a$ $SE$ $\lambda^b$</td>
<td>Odds ratio $\Lambda^a$ $SE$ $\lambda^b$</td>
</tr>
<tr>
<td><strong>Interview structure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective-job related</td>
<td>1.07 0.45 0.41** 2.93</td>
<td>0.90 0.56 0.35* 2.46</td>
</tr>
<tr>
<td>Standardized administration</td>
<td>0.18 0.24 0.12 1.20</td>
<td>0.40 0.34 0.25 1.49</td>
</tr>
<tr>
<td>Multiple interviewers</td>
<td>0.15 0.27 0.08 1.17</td>
<td>0.40 0.36 0.20 1.50</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filed under Title VII</td>
<td>-2.75 1.51 -0.38* 0.06</td>
<td></td>
</tr>
<tr>
<td>Number of plaintiffs</td>
<td>-0.31 0.40 -0.15 0.74</td>
<td></td>
</tr>
<tr>
<td>Unskilled jobs</td>
<td>-0.30 0.73 -0.08 0.74</td>
<td></td>
</tr>
<tr>
<td>Managerial jobs</td>
<td>0.45 0.82 0.11 1.57</td>
<td></td>
</tr>
<tr>
<td>New hire decision</td>
<td>0.45 0.82 0.11 1.57</td>
<td></td>
</tr>
<tr>
<td>Judge political appointment</td>
<td>1.69 0.68 0.46** 5.43</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Blank spaces = no information obtained. $\chi^2 (3, N = 84) = 13.88, p < .05$ (Model 1). $\chi^2 (9, N = 84) = 33.01, p < .05$ (Model 2).

---

**Table 5**

*Logistic Regression Analysis of Verdict on Interview Structure for Disparate-Impact Claims*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model 1 (Structure only)</th>
<th>Model 2 (Controls added)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Lambda^a$ $SE$ $\lambda^b$</td>
<td>Odds ratio $\Lambda^a$ $SE$ $\lambda^b$</td>
</tr>
<tr>
<td><strong>Interview structure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective-job related</td>
<td>2.39 0.84 0.96** 10.96</td>
<td>1.75 1.10 0.70* 5.79</td>
</tr>
<tr>
<td>Standardized administration</td>
<td>-0.31 0.45 -0.19 0.74</td>
<td>0.21 0.60 0.13 1.24</td>
</tr>
<tr>
<td>Multiple interviewers</td>
<td>-0.18 0.41 -0.10 0.85</td>
<td>-0.13 0.52 -0.07 0.88</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of plaintiffs</td>
<td>-0.58 0.59 -0.29 0.56</td>
<td></td>
</tr>
<tr>
<td>Promotional decision</td>
<td>-1.14 1.11 -0.29 0.32</td>
<td></td>
</tr>
<tr>
<td>Private–nonunion organization</td>
<td>-1.10 0.98 -0.28 0.37</td>
<td></td>
</tr>
<tr>
<td>Public (other) organization</td>
<td>1.12 1.40 0.27 3.08</td>
<td></td>
</tr>
<tr>
<td>Judge political appointment</td>
<td>1.72 0.99 0.48* 5.59</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Blank spaces = no information obtained. $\chi^2 (3, N = 46) = 14.75, p < .05$ (Model 1). $\chi^2 (8, N = 46) = 23.61, p < .05$ (Model 2).

---

*a* Unstandardized, *b* Standardized, *c* Democrat = 0, Republican = 1.

*p < .10, **p < .05.*
claims decided for the plaintiff tended to involve Title VII filings, multiple plaintiffs or class actions, unskilled jobs, nonmanagerial jobs, new-hire decisions, job-assignment decisions, and democratic judges. Disparate-impact claims decided for the plaintiff tended to involve multiple plaintiffs or class actions, promotional decisions, training decisions, private-nonunionized organizations, and democratic judges.

The previous logistic regression analyses were repeated with significant descriptors included as control variables. One variable was excluded from each analysis (job-assignment decisions for disparate-treatment claims and training decisions for disparate-impact claims) because their variances were so low that the analysis could not iterate to convergence. As shown in Tables 4 and 5 (Model 2), the results were unchanged for both types of claims. The objective-job related composite was still significant, and it was still the only composite that was significant. Therefore, the inclusion of control variables had no meaningful effect on the results. It is noteworthy that the judge's political appointment was the only control variable that was significant in both analyses. Its odds ratio was also quite large, indicating the potential importance of considering such personal factors when studying court case outcomes.

Discussion

Summary of Results

Past research has shown that structuring the interview can be a key to enhancing its reliability and validity. The purpose of this study was to test the thesis that structuring can also be a key to enhancing its legal defensibility. Two approaches were taken to test this thesis, one conceptual and one empirical. Both were strongly supported.

Conceptually, interview structure may be linked to litigation outcomes through a variety of mechanisms. Standardizing the administration of the interview increases consistency across candidates, and enhancing job relatedness helps justify business necessity. Structuring may also reduce the overall subjectivity and, thus, potential for bias in the decision-making process, and many aspects of structure are likely to enhance perceptions of procedural justice.

Empirically, the many aspects of interview structure reduced to three composites. The first consisted of interview objectivity and job relatedness, such as objective and specific criteria, trained interviewers who are familiar with job requirements, and validation evidence. The second composite consisted of standardized administration, such as guidelines, minimal discretion, common questions, and consistency. The final composite consisted of multiple interviewers, such as panels and reviews of interviewer decisions.

Most items (Tables 1 and 2) and most composites (Table 3) were found to be related to verdicts for the defendant in both disparate-treatment and -impact claims. When multivariate analyses considered all aspects of structure simultaneously and in the presence of control variables, the objective and job-related composite was most important. These aspects of structure may be more apparent to judges who are attuned to the legal significance of job relatedness than the finer nuances of standardized administration emphasized by researchers.

It was anticipated that standardized administration would be especially important to disparate-treatment claims and job relatedness would be especially important to disparate-impact claims. However, despite the fact that the formal burdens of proof differ, the important aspects of structure showed remarkable similarity between types of claims.

The structure measures were intercorrelated, suggesting that interviews well developed on some aspects are also well developed on others. This should come as no surprise because improvements in interviews often represent combinations of many aspects of structure (e.g., M. A. Campion, Campion, & Hudson, 1994; Janz, 1982; Latham, Saari, Pursell, & Campion, 1980; Motowidlo et al., 1992). Also, despite relative differences in the strengths of the results, all three categories of structure are important.

Recommendations for Practice and Research

The most conservative conclusion from the empirical findings is that employers who used structured interviewing in the past, and successfully rebutted employment discrimination cases, probably saw aspects of structure that the judge relied on to justify the verdict. A stronger conclusion is that the empirical findings together with the conceptual analysis suggest that structuring the interview might enhance the ability to withstand legal challenge. Prior evidence on the superior validity of structured interviews, and the potentially enhanced legal defensibility of such interviews described in this study, lead to the practical recommendation to use the aspects of structure shown in Tables 1 and 2.

This study may hold recommendations for expert witnesses who testify in court cases on the interview. If judges attend to interview structure in making their decisions, it may be advisable to focus testimony (either for or against) on these aspects. In addition, focus should be more on objective and job-related aspects of structure, rather than on standardized administration and multiple interviewers, because this reflects the emphasis of the judges in this study.

Judges' opinions may not actually identify causes of bias in the interview process. This is a topic for future research. For example, it is not clear that job-related and
objective aspects of interviews reduce bias more than standardized administration, either in terms of the psychometric definition of bias (i.e., test fairness or equality of prediction) or the lay definition of bias (i.e., adverse impact). Research might also focus on how structure influences the psychological and cognitive processes related to bias (e.g., perhaps some aspects of structure help prevent the formation or operation of stereotypes).

Limitations

A number of factors limit the potential generalizability of the findings (Roehling, 1993). For example, cases that go to trial may differ from those that do not. The findings of this study cannot prove which cases are likely to go to court. Furthermore, court cases are written up after the decisions are made. They are meant to explain and justify the decision. As such, there is concern that some aspects of the interview or situation were not included, as well as a more generalized concern over the accuracy of post hoc, self-reported explanations of behavior.

The present study focused on decisions made by judges in nonjury or bench trials. In such cases, the judge decides issues of law and issues of fact. The Civil Rights Act of 1991 amended Title VII to give the parties the right to demand a jury trial where intentional discrimination is alleged and the plaintiff seeks compensatory or punitive damages. In jury trials, the judge still decides issues of law, but the jury decides the issues of fact. Thus, there is a question regarding the extent to which the decisions of district court judges remain important and provide useful information. It may be argued that the findings are valuable for several reasons.

First, compensatory and punitive damages are not available in cases involving only disparate impact. There is no jury in those cases. Judges decide issues of law and fact, and so the present findings remain directly relevant. Second, even in jury trials, judges are often called upon to make factual determinations when deciding motions for summary judgment and directed verdicts. Although the standard applied is a higher one, they are asked to consider the same underlying factual issues. Third, whereas parties have a right to request a jury trial, the parties may elect not to exercise that right. Thus, there will continue to be some Title VII disparate-treatment cases in which the judge decides issues of law and fact. Fourth, decisions in trials are likely to be reasonable indicators of the outcomes in arbitration cases involving employment discrimination claims. Arbitrators, like trial judges, decide issues of law and issues of fact. Furthermore, the demand for guidance in arbitration cases involving employment discrimination can be expected to increase given that federal courts have embraced arbitration as a means for relieving backlogged courts, and employers are responding by establishing more employment arbitration procedures (Bales & Burch, 1994).

Fifth, juries’ deliberations will be subject to instructions from judges as to what factors to consider in deciding issues of fact, and jury decisions are subject to review by the judges on the basis of standards of reasonableness. As such, judges might predispose juries to consider aspects of interview structure. Sixth, it is likely that lawyers will argue partly on the basis of precedents from previous cases, even if those cases were decided by judges. As such, the importance of interview structure is likely to be brought forward to the juries. Seventh, the same aspects of structure that appear to prevent discrimination in the eyes of judges are likely to be similar in the eyes of jurors because they are intuitively logical (e.g., standardized administration is logically related to preventing disparate treatment). Finally, this study somewhat mitigated bias associated with this kind of self-report data by controlling for the most widely recognized potential source of “extra-legal” systematic bias—the political affiliation of the judge—and by incorporating other relevant control variables (e.g., type of job, nature of employment decision, etc.).

Conclusion

It is gratifying that the factors judges consider when explaining court decisions correspond to recommendations long made by industrial and organizational psychologists for improving the interview. It is reassuring to see that those aspects of interviewing systems found through research to contribute to enhanced reliability and validity are also viewed by the courts as enabling the interview to protect against unlawful employment discrimination.

References


Reynolds v. Sheet Metal Workers Local 102, 22 EPD P30,739 (1980).


THE EMPLOYMENT INTERVIEW ON TRIAL


Appendix

Case Citation List

Multiple sources are listed for certain cases. EPD (Employment Practices Decisions) cases published by Commerce Clearing House. FEP (Fair Employment Practices) cases published by Bureau of National Affairs.


Colon v. Sorensen, 668 F. Supp. 1319 (Nebraska); 45 FEP Cases 478; 45 EPD P37,646 (1987).


Gay v. Waiters’ and Dairy Lunchmen’s Union, Local No. 9, 489 F. Supp. 282 (N.D. Calif.); 22 FEP Cases 281; 23 EPD P30,928 (1980).

Pate v. Alameda-Contra Transit District, 21 FEP Cases 1228 (1979).
Peltier v. City of Fargo, 396 F. Supp. 710 (North Dakota); 10 FEP Cases 1452 (1975).
Reynolds v. Sheet Metal Workers Local 102, 22 EPD P30,739 (1980).

Revision received September 25, 1996
Accepted July 14, 1997