PERSONNEL PSYCHOLOGY
1997, 50

# EXPLAINING U.S. COURTS OF APPEALS DECISIONS INVOLVING PERFORMANCE APPRAISAL: ACCURACY, FAIRNESS, AND VALIDATION

JON M. WERNER, MARK C. BOLINO
Department of Management
University of South Carolina

Accuracy and due process perspectives were used to extend policy-capturing research concerning employment discrimination case law. Two-hundred ninety-five usable U.S. Circuit Court decisions concerning performance appraisal were located from 1980–1995. In both chi-square and multivariate LOGIT analyses, decisions were explained by: use of job analysis, provision of written instructions, employee review of results, and agreement among raters. Contrary to hypotheses, appraisal frequency and type (traits vs. behaviors or results) were unrelated to judicial decision. Rater training approached significance in chi-square analysis. Of other variables checked (e.g., type of discrimination claim, statutory basis, class action status, year of decision, circuit court, type of organization, purpose of appraisal, evaluator race and sex), only circuit court approached significance. We conclude that issues relevant to fairness and due process were most salient to judicial decisions; issues pertaining to accuracy were important, yet validation was virtually ignored in this sample of cases.

Today there is no dispute that performance appraisal practices are subject to employment legislation such as Title VII of the 1964 Civil Rights Act. Further, many researchers and practitioners view performance appraisal as an employment "test" covered by the *Uniform Guidelines on Employee Selection Procedures* (1978). The *Guidelines* were adopted in 1978 by the Equal Employment Opportunity Commission and other regulatory agencies, and emphasize the need for employers to validate *all* employment criteria, both "objective" and "subjective," where adverse impact has been found. This is the official position of

We would like to thank John Hollenbeck, Sandra Gleason, Hoyt Wheeler, Brian Klaas, Martin Malin, John Grego, Samantha Wolf, Tom Ruprecht, Bob Strack, Scott Keyes, and the anonymous reviewers for their valuable assistance at various stages of this research. An earlier effort at similar research was presented at the National Academy of Management Convention in San Francisco, August, 1990.

Correspondence and requests for reprints, or a list of Courts of Appeals cases used in this analysis, may be obtained from Jon M. Werner at the Department of Management, University of South Carolina, Columbia, SC 29208 or e-mail wernerj@darla.badm.sc.edu.

Division 14 of the American Psychological Association as well (*Amicus Curiae Brief,* 1988).

Initially, however, it was unclear whether the courts would view performance appraisal as an employment test (Schneier, 1978). Acting on the assumption that appraisals should be viewed in this manner, Feild and Holley (1982) utilized quantitative analyses to determine which variables most influenced judicial decisions in employment discrimination cases. They studied legal cases from 1965 to 1980 where performance appraisal was an issue in an employment discrimination charge. Sixty-six cases were located, including two state court, 46 federal district court, 16 court of appeals, and two U.S. Supreme Court decisions. Five variables distinguished between decisions for plaintiffs and decisions for defendants: use of job analysis, behavior-oriented appraisal systems, provision of written instructions for evaluators, review of appraisal results with employees, and type of organization (industrial vs. non-industrial). When the legal ramifications of performance appraisal are discussed, recommendations similar to their findings are often made (Barrett & Kernan, 1987; Martin & Bartol, 1991; Milkovich & Boudreau, 1994; Veglahn, 1993).

Recently, the strong measurement emphasis in performance appraisal has been likened to a "test metaphor" (Folger, Konovsky, & Cropanzano, 1992). That is, the primary goal of performance appraisal has been viewed as rating accuracy. Much research has emphasized the psychometric properties of rating formats, and the training of raters to reduce bias (viewed as inaccuracy; cf. Landy & Farr, 1980). Folger et al. (1992) criticized this approach on three grounds. First, the rapidly changing nature of work makes it increasingly difficult to obtain reliable and valid performance measures under the best of circumstances. Second, accuracy is further reduced by rater limitations. Cognitive limitations have been extensively documented in the decision making and information processing literatures (Bazerman, 1990; Ilgen, Barnes-Farrell, & McKellen, 1993). Rater motivation to rate accurately has also been questioned (Longenecker, Gioia, & Sims, 1987). Third, a test metaphor assumes the existence of a unitary performance criterion. This minimizes or ignores differences in values and goals between raters, ratees, and other stakeholders (Tsui, 1990), and by default, establishes organizational goals as the accuracy criterion.

As an alternative, Folger et al. (1992) proposed a due process framework. In the legal arena, there are at least three essential features of due process. Before a decision is made, both parties must have adequate notice, a fair hearing, and the expectation that judgments will be based on evidence (Forkosch, 1958). Legal proceedings more often revolve around "rendering justice" (i.e., satisfactory conflict resolution)

rather than "determining truth" (i.e., resolving disputes about the facts in a given case). Thibaut and Walker (1978, p. 556) argued that "the primary goal of the legal process must be the attainment of distributive justice between the parties rather than the realization of the most accurate view of reality." This emphasis is obviously quite different from the traditional emphasis on accuracy in most performance appraisal research. For example, the use of job analysis, behavior-based appraisals, and validation is typically recommended as a means of reducing bias and increasing test or appraisal accuracy (Feild & Holley, 1982; Veglahn, 1993). However, the arguments from Thibaut and Walker would suggest that courts of law may be more likely to view appraisal characteristics as important insofar as they impact the due process or fairness of the proceedings. As described in more detail below, the due process literature has recommended practices such as employee review of appraisal results and providing more frequent appraisals as important for increasing the perceived fairness of the performance appraisal process (Folger et al., 1992). Many variables, including rater training or interrater agreement are likely to be salient to both the fairness and the accuracy of appraisals, but even here, their impact on judicial decisions may be as much for their perceived relationship to appraisal fairness as for their expected effect on appraisal accuracy.

In order for participants to perceive justice in legal proceedings, the procedures used must be viewed as fair (Folger & Konovsky, 1989; Lind & Tyler, 1988). Folger et al. (1992) suggested that procedural justice research in psychology and law (see Lind & Tyler, 1988) be applied to performance appraisal. Toward this end, Taylor, Tracy, Renard, Harrison, and Carroll (1995) reported that employees whose managers had been trained to emphasize due process in their appraisals had more favorable reactions to the appraisal process than did employees whose managers received no training, even though performance ratings in the experimental condition were lower than those received by control condition subjects.

Because of the importance of legal challenges to performance appraisal, the primary purpose of our study is to replicate and extend policy-capturing research on this topic. Concerning replication, we expect generally similar findings to those obtained by Feild and Holley (1982). We extend this line of research in two primary ways. First, research of this type has recently been critiqued (Roehling, 1993), and we respond to Roehling's recommendations, so as to produce results that are legally and methodologically sound. Second, employee lawsuits can be seen as a "strong" form of negative employee reaction (Taylor et al., 1995). The due process framework just described (Lind & Tyler, 1988;

Thibaut & Walker, 1978) has recently been applied to performance appraisal (Folger et al., 1992; Korsgaard, Roberson, & Rymph, 1996; Taylor et al., 1995). In this study we use both test and due process perspectives to understand performance appraisal court decisions. It is clearly not so simple as to say that some of our variables are primarily related to accuracy, and others to fairness. Thus, we derived our hypotheses using both perspectives. We next review the literature to highlight weaknesses in this line of research and to describe our handling of these issues in the current study.

*Literature Review*

As Feild and Holley (1982) noted, their sample of 66 cases was very small for the generalizations made. Variables not discussed in the legal opinions were coded as missing, with some analyses based on sample sizes as small as 12. This greatly weakens the power of their study to find statistically significant differences when they do in fact exist. A larger sample is imperative in order to verify both significant and nonsignificant findings from Feild and Holley. Although related articles have been published since 1982, none adequately tested their results in "new or larger samples of discrimination cases as they appear in the courts" (Feild & Holley, 1982, p. 399). Feild and Thompson (1984) replicated four of the significant findings from Feild and Holley (1982) using federal district and court of appeals cases from 1980 to 1983 (type of organization was not significant). However, their study included only 31 cases, and variables such as rater training were not measured, even though identified by prior reviews as important (Bernardin, Beatty, & Jensen, 1980).

Some subsequent articles have provided excellent discussions of the legal aspects of performance appraisal (Ashe & McRae, 1985; Barrett & Kernan, 1987; Martin & Bartol, 1991; Martin, Bartol, & Levine, 1986; Veglahn, 1993). Martin and colleagues studied court cases relating to four purposes: promotion, discharge, layoff, and merit pay (Martin & Bartol, 1991; Martin et al. 1986). Barrett and Kernan (1987) reviewed performance appraisal court cases dealing with terminations only since *Brito v. Zia* (1973). Each of these articles utilized a narrative/legal review of recent court cases. Although this approach is valuable, the use of statistical analyses, as in the present study, can provide information beyond that available from narrative reviews.

For example, some research has focused specifically on age discrimination. Schuster and Miller (1981) and Faley, Kleiman, and Lengnick-Hall (1984) narratively reviewed recent court decisions, while Schuster and Miller (1984) and Miller, Kaspin, and Schuster (1990) empirically

analyzed such cases. An interesting point made by Faley et al. (1984) is that "validity concerns" appear less important in cases brought under the Age Discrimination in Employment Act (ADEA) than in Title VII cases. In support of this, Miller et al. (1990) documented how little validation issues are discussed in ADEA court cases. By using multiple bases for discrimination complaints, the current research can, among other things, test whether courts treat age discrimination cases differently than other cases.

In an important review, Roehling (1993) criticized policy-capturing studies of judicial opinions on a number of grounds. Three of Roehling's (1993) concerns (changes in the law over time, sample bias, and data aggregation problems) can largely be addressed via methodological means, and are addressed below in the methods section. However, two of the "dangers" he highlighted (using data drawn from judicial opinions, and using statistical analyses to study legal issues) strike at the essence of this form of policy-capturing research, and are addressed directly. The first argument raised by Roehling (1993) is that judicial opinions are written to justify a particular decision, and thus may not capture all the information actually used to reach such decisions. The concern here is that policy-capturing research draws its data from the information contained in written judicial opinions, and is thus constrained by what judges choose to include in such opinions. Despite this acknowledged limitation, there is value in ascertaining the extent to which judges make note of issues such as use of job analysis and providing raters with written instructions as "official" reasons for their decisions. The reason for this is that such factors are generally recommended as "best practices" by human resource management scholars (Barrett & Kernan, 1987; Milkovich & Boudreau, 1994; Veglahn, 1993), and have been codified in documents such as the *Uniform Guidelines* (1978). In this study, we tested whether judges used these factors to justify their decisions, while at the same time testing for other factors that may have influenced their decisions (such as rater race and sex, geographic location, and court circuit).

Another of Roehling's (1993) criticisms was that policy-capturing studies have used statistical analyses that were either too simple or too complex. That is, traditional legal analysis is sufficient to determine whether defendants have provided "legitimate nondiscriminatory reasons" for their actions (*Texas Department of Community Affairs v. Burdine,* 1981). However, for other factors, interactions may be present. Roehling (1993, p. 495) recommended that future research test for interactions between type of claim (race, sex, age) and performance appraisal characteristics (use of job analysis, behavior-based appraisals, etc.). In response, we used logistic regression analyses to test for the main effects

of variables such as type of claim, and also for potential interactions with our selected performance appraisal characteristics.

*Hypotheses*

Policy-capturing research can indicate which variables influence decisions in a given area (Nagel & Neef, 1979). A major question addressed by this study is whether courts hold appraisal systems to the standards found in the *Uniform Guidelines* (1978). Two sets of hypotheses were made. The first set of hypotheses should replicate Feild and Holley (1982). That is, appraisals should: (a) be based on job analyses, (b) emphasize behaviors or results, rather than traits, (c) include specific written instructions to raters, and (d) allow review of appraisal results by employees. These variables continue to be recommended in legal reviews of performance appraisal (Ledvinka & Scarpello, 1991), and should increase both the accuracy and fairness of appraisals. Changes in the law should not have made them any less important from 1980–1995 than prior to 1980. In fact, increased awareness of legal issues and changes in the law might lead these variables to be even more important in the latter time period. Although Feild and Holley (1982) found that type of organization was significantly related to judicial decisions, this variable was not expected to explain court decisions in the current study, since this was likely an artifact of the enforcement of Title VII for governmental organizations prior to 1980 (Feild & Thompson, 1984).

*Hypothesis 1:* Decisions in favor of organizations will be more likely when the appraisal system is based on a job analysis.

*Hypothesis 2:* Decisions in favor of organizations will be more likely when the apraisal is behavior- or results-oriented.

*Hypothesis 3:* Decisions in favor of organizations will be more likely when specific, written instructions are given to evaluators.

*Hypothesis 4:* Decisions in favor of organizations will be more likely when appraisal results are reviewed with employees.

Although the presence of information concerning appraisal validity and reliability did not explain judicial decisions in Feild and Holley (1982), courts might be expected to pay increased attention to validity and reliability, because performance appraisal is often viewed as subject to the *Uniform Guidelines* (1978). For example, in *Watson v. Fort Worth Bank & Trust* (1988), Justice O'Connor accepted a promotion decision based on subjective criteria as falling under Title VII coverage (and thus amenable to disparate impact analysis). However, formal validation studies are most likely to be included in legal opinions concerning *disparate impact;* yet, these make up only the minority of employment

discrimination court cases (Roehling, 1993). Further, validation makes sense for appraisal ratings only when such information is used to predict something (such as a promotion or transfer decision; Barrett & Kernan, 1987). Thus, in the current study, we looked for any mention of validation efforts, but because of the expected paucity of data, no formal hypotheses were made in this regard. Similarly, discussions of reliability (in the psychometric sense) were not expected in written court decisions. However, a hypothesis was made concerning a more informal type of interrater agreement. This is discussed next.

The second set of hypotheses concern variables that Feild and Holley (1982) predicted would influence judicial decisions, but which did not attain statistical significance in their study. On both theoretical and methodological grounds, we argue that these variables should be related to court outcome. The theoretical rationale is that agreement among multiple raters, evaluator training, and increased rating frequency should increase both the accuracy and the perceived fairness of the appraisal system. The use of multiple raters is widely recommended (Conway, 1996; Ledvinka & Scarpello, 1991; Veglahn, 1993), and judges would be expected to pay particular attention to corroborative evidence when settling employment discrimination disputes (Thibaut & Walker, 1978). We labelled this variable "triangulation," as Folger et al. (1992) suggested that triangulating on truth from multiple directions should increase the perceived fairness of the appraisal process. Similarly, past research has found that rater training can increase both the accuracy (Cardy & Keefe, 1994; Pulakos, 1984; 1986; Sulsky & Day, 1994) and the perceived fairness (Korsgaard et al., 1996; Taylor et al., 1995) of the appraisal process. Finally, more frequent appraisals are more likely to be accurate, and have also been linked to greater employee satisfaction with the evaluation process (Landy, Barnes, & Murphy, 1978). Folger et al. (1992, p. 142) described "feedback given on a regularly recurring and timely basis" as critical to adequate notice, and hence to fairness perceptions. The methodological grounds for retesting these hypotheses was that, in Feild and Holley (1982), small sample size was an acute problem for each of these variables.

*Hypotheses 5:* Decisions in favor of defendants will be more likely when there is triangulation among multiple raters.

*Hypothesis 6:* Decisions in favor of defendants will be more likely when evaluators receive training.

*Hypothesis 7:* Decisions in favor of defendants will be more likely when formal evaluations are conducted more frequently.

## *Method*

### *Selection of Legal Cases*

The WESTLAW computer data base was searched for the years 1980–1995. West Publishing Company publishes the Federal Reporter for U.S. Courts of Appeals decisions, and the Federal Supplement for U.S. District Court decisions. Reynolds and Richman (1981) estimated that these series publish approximately 38% of all Appeals decisions and 10–15% of District Court decisions (the Courts of Appeals decide what gets published based on perceived importance and precedential value, Cohen & Berring, 1983). All published decisions since 1980 are included on WESTLAW. Thus, although this data base does not cover all Courts of Appeals decisions for the time period, most important and relevant cases should be included.

Queries were made of the Supreme Court, Courts of Appeals, and District Court data bases on WESTLAW. All cases from January, 1980 until June, 1995 were sought where performance appraisal or performance evaluation was mentioned, as well as the terms "employment" and all forms of the word "discriminate." Only decisions at the Courts of Appeals level were included in this study. Feild and Holley (1982) used mainly District Court decisions; thus, the datasets are not strictly comparable. Appeals Court decisions were chosen because: (a) they have greater legal precedence than District Court decisions, (b) differences in patterns of court outcomes can be tested across 13 court of appeals circuits, unlike the far larger number of federal district courts, (c) all located cases for this period could be read and considered for inclusion; due to practical constraints, this was not possible with the 1,870 district court decisions turned up by the search (as noted by Roehling, 1993, year of decision may be relevant to this study), and (d) aggregating across trial and appellate levels has been criticized for ignoring potential differences between levels (Roehling, 1993).

A total of 627 possible Court of Appeals decisions were located from 1980–1995. Of these, 308 (49%) could be included in the data analysis. Many citations were duplicates or were to cases not published. The largest number of discarded cases were those where performance appraisal was either not central to the case (e.g., a hiring case), or there was another basis of complaint than listed above (e.g., an unfair labor practice).

After the initial data analysis, 25 of the 308 cases had missing data for judicial decision. Most of these were cases that had been vacated or remanded back to district courts. In some decisions to vacate or remand, the judicial opinion provided guidance concerning how to code

judicial decision, (i.e., the act of vacating or remanding the case could be classified as a "victory" for one of the parties to the dispute.) In these 25 cases, however, decision could not be determined. Using Auto-Cite on LEXIS, plus numerous phone calls to lawyers' and court clerks' offices, 12 of the 25 missing decisions were subsequently coded, leaving a final sample of 295 for all analyses.

Of these 295 cases, 109 (35.4%) concerned age discrimination, 102 (33.1%) involved charges of race discrimination, and 50 (16.2%) dealt with sex discrimination. Thirty-three cases (10.7%) dealt with charges based on national origin, creed, handicap, or involved multiple bases (e.g., sex and race). One case could not be coded for basis of discrimination charge. 49.3% of the cases were filed under Title VII alone, 34.7% under the ADEA alone, 4.5% were filed as constitutional claims, and 11.5% of the cases involved a combination of statutory bases.

*Variables*

The dependent variable in this study was the decision reached by the court. A decision in favor of the plaintiff was coded 0; a decision in favor of the defendant was coded 1. Cases that were remanded to lower courts without opinion were not included in our sample. For cases that could not be clearly coded as "for plaintiff" or "for defendant" (Roehling, 1993), the following decision rule was used: If any significant portion of the ruling was in favor of the plaintiff(s), the decision was coded 0. Although this rule favors plaintiffs, we feel it is consistent with how such decisions are viewed by participants in the legal system, as well as with the wording of our hypotheses.

Categorical codes were used for the independent variables of interest. For example, use of job analysis to develop appraisal (and other variables with a yes/no response) were coded no = 0, yes = 1. Similar codes were developed for appraisal type (trait = 0, behavior- or results-oriented = 1); evaluators given specific, written instructions on how to complete appraisals, results of appraisals reviewed with employees, triangulation among raters, and rater training. Frequency of appraisal was coded as follows: 0 = less than once a year, to 4 = more often than every 3 months.

*Control variables.* A number of variables were included to check for their potential influence on judicial decisions. Variables that might be related to court outcome include organization type, race and sex of evaluator(s), the geographic location where the suit is filed, and appraisal purpose. These variables were measured to see if statistical power was an issue in the nonsignificant findings obtained by Feild and Holley

(1982). However, no hypotheses were made concerning the relationships between these variables and judicial decision.

Type of organization was coded: private = 0; public = 1. Race of evaluator was coded: non-minority/white = 0, minority or mixed = 1; as was sex of evaluator: male = 0, female or a mixed group of evaluators = 1. Geographic location was coded with three dummy variables for west, midwest, and east, with south as reference group; and was also tested dichotomously, where non-south = 0, and south = 1. Similarly, purpose of appraisal was coded in two ways (for separate analyses): using three dummy variables (for layoff, transfer, and discharge, with promotion as the reference group); and dichotomously, (i.e, purposes other than promotion = 0; promotion = 1.) The number of evaluators used was also noted.

Five variables were included to address issues raised by Roehling (1993). First, the evolving nature of the law poses particular challenges for data analysis. For this reason, decision year was recorded and used in various aggregation schemes to test for changes over time periods (e.g., the Civil Rights Act of 1991 opened the possibilities of jury trials and punitive damages to a wider number of discrimination claims, and we tested for differences between cases filed before and after this act took effect). Second, because different legal requirements apply when cases are filed under disparate impact versus disparate treatment theories of discrimination (Ledvinka & Scarpello, 1991), we coded the legal theory under which each case was filed, to test for differences between them, and to ensure their comparability before aggregating for data analyses. Third, cases were coded for whether they had been brought by individuals or were class actions. Fourth, because different bases for discrimination charges could potentially influence decisions, the basis for the discrimination charge was recorded (race, sex, age, disability, etc.). These were then coded using three dummy variables for race, sex, and age, with "other" as the reference group (including disability, creed, national origin, and multiple bases).

Finally, Roehling (1993) argued that environmental factors can influence the particular disputes being decided by the courts. Political science research has found a moderate but clear influence of party affiliation on court decision making (Goldman, 1975; i.e., whether the judicial nomination originated from the Republican or the Democratic party). Further, some circuits are considered politically liberal, while others are more conservative (Spaeth, 1985). We tested whether differences would emerge between circuits in the present sample as well (cf. Schuster & Miller, 1984). The circuit where the decision was rendered was recorded. Although this does not consider differences among judges, finer breakdowns were not feasible.

Variables were coded for whether information had been presented concerning the reliability or validity of the appraisal system. As a final check, we coded whether defendants met their evidentiary burden of proof. For disparate treatment cases, the three-step procedure from *Texas Department of Community Affairs v. Burdine* (1981) and *McDonnell-Douglas v. Green* (1973) was used to make this determination. After the plaintiff establishes his or her initial burden, the defendant then has a rebuttal burden (i.e., to articulate a legitimate nondiscriminatory reason for its action.) Plaintiffs then have an opportunity to challenge these arguments. The burdens of proof are different in disparate impact cases (Ledvinka & Scarpello, 1991). However, the plaintiff/defendant/plaintiff pattern is also followed in disparate impact cases, and this shifting burden of proof was similarly coded in disparate impact cases.

*Procedure*

Information on the case characteristics was obtained from the judges' written decisions. Related decisions from other courts were located whenever possible, primarily at the district court level, although a few Supreme Court decisions were also found (e.g., *Bazemore,* 1986; *Watson,* 1988). In an effort to reduce the amount of missing data, these decisions were read as well, and any additional information concerning the independent variables was added to our dataset.

This rating scheme was used by five raters to code all the cases used in this study. Cases from 1980–1988 were coded by the first author. From this initial sample, 10 cases were randomly selected (from a pool of 120) and were independently coded by a rater unaffiliated with the study. This produced measures of interrater agreement ranging from 0.87 to 1.00, with a mean reliability of 0.94. A similar procedure was followed for three raters who coded cases from 1988–1990, 1990–1992, and from 1992–1995. None of the raters began coding cases in their set until the level of interrater agreement between their ratings and the ratings given by the first author was at least 0.95.

*Data Analysis*

Univariate chi-square analyses were used to test six of our hypotheses. Because frequency was treated as a continuous variable, this was analyzed using univariate logistic regression. A power analysis revealed that 84 cases would be needed for univariate analyses to have an 0.80 probability of detecting a moderate population effect size ($r = .30$), using an alpha of $p = .05$ (Cohen & Cohen, 1983). In order to examine

the effects of the independent variables while controlling for common variation, a multivariate logistic regression analysis (LOGIT) was also performed. Six of our control variables were also included in this analysis. LOGIT was chosen because decision was coded dichotomously. LOGIT models predict the likelihood for a particular category of a dichotomous variable. In this case, we were predicting the likelihood of a decision in favor of the organization. Mean substitution was used for missing values, as listwise and pairwise deletion of cases was infeasible. Donner (1982) found that mean substitution is relatively effective for estimating the coefficients of variables with missing data when the correlations among these variables are weak, even when the proportion of missing cases is fairly high. Donner does not provide precise guidelines for what is meant by weak intercorrelations or a high amount of missing data. However, statistical simulations we ran suggested that beta estimates would be biased not more than 5% even when intercorrelations among the independent variables were 0.50 (a large correlation, according to Cohen & Cohen, 1983), and the proportion of missing data was as high as 90%. Further, in our simulations, the biasing effect was more strongly related to the intercorrelations among the independent variables than to the amount of missing data.

## *Results*

### *Judicial Decision*

Overall, 58.6% of the cases in our sample were decided in favor of defendants. Coding the dependant variable was straightforward in the vast majority of cases. Approximately 20% of the cases in our sample were split decisions, where portions of the decision favored plaintiffs, and other portions favored the defendants. Yet, analyzing the content of these split decisions made coding clear cut in roughly 75% of these cases. Thus, our decision rule (if any significant portion of the ruling favors the plaintiff, code as zero) needed to be invoked in 15 (or 5%) of the cases.

### *Control Variables*

Analyses were conducted to ensure the appropriateness of our aggregation methods (Roehling, 1993). First, a chi-square analysis revealed that year of decision was unrelated to decision ($\chi^2 = 15.36, p = .43$). To further justify our aggregation of cases across the time period, the effect of year was then tested using several aggregating schemes. For example, tests were conducted to determine whether cases after November, 1991

(i.e., cases potentially affected by the Civil Rights Act of 1991) differed significantly from others in the set with respect to judicial decision. None of the aggregation combinations revealed significant differences. As a last check concerning the potential impact of year of decision, we added this variable into logistic regressions, as well as a variable capturing the interaction between decision year and each independent variable. Neither year nor any of these interaction terms was statistically significant.

Whether a case was a class action or not was not significantly related to judicial decision ($\chi^2 = 1.68, p = .19$), and class action status did not interact with any independent variable in explaining judicial decisions. The use of a disparate impact versus disparate treatment framework was not related to judicial decision ($\chi^2 = 9.33, p = .32$). In fact, the analyses reported below were unchanged when disparate impact cases were dropped from the analyses (such cases constituted 7.1% of the current sample). Tests for interactions among the independent variables and type of claim revealed one significant interaction effect, namely between the use of job analysis and type of claim ($p = .046$). An inspection of the cell means indicated that, similar to the main effect, decisions in favor of defendants were more likely in disparate treatment cases where job analysis had been used. However, in the six disparate impact cases where job analysis was mentioned, results were opposite of our expectations, (i.e., only one of the five cases where job analysis was present was won by the defendant, whereas in the one case where job analysis was not present, the decision was also in favor of the defendant.)

The basis for the discrimination claim (age, race, sex, etc.) had no effect on decision ($\chi^2 = 1.89, p = .60$), nor did statutory basis for the charge ($\chi^2 = .63, p = .89$). There were no interactions between any of the independent variables and either basis for the discrimination charge or statutory basis. Consistent with Feild and Holley (1982), the following variables were unrelated to judicial decision: appraisal purpose, geographic location, and race and sex of evaluator. Further, none of these variables interacted with any of the independent variables in explaining judicial decision. Analyses were unaffected by the two methods of coding purpose and geographic location.

Some differences were observed in the pattern of decisions by circuit. The highest percentage of cases decided for defendants were in circuits 1, 4, and 5 (88%, 68%, and 71%, respectively), with the lowest percentage of cases decided for defendants in the Washington, D.C. circuit (44%). A logistic regression analysis using 11 dummy variables (combining the District of Columbia and Federal Circuits) approached significance ($\chi^2 = 19.17, p = .06$). As expected (based on Roehling, 1993), validation and reliability (in the psychometric sense) were mentioned fewer than 10 times each, and thus could not be used for data analyses.

TABLE 1

*Correlation Matrix*[a]

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. Job analysis | 49 | | | | | | | |
| 2. Appraisal type | .31* | 167 | | | | | | |
| 3. Instructions | .93** | .25 | 50 | | | | | |
| 4. Employee review | .30 | .26** | .25 | 156 | | | | |
| 5. Triangulation | .32 | −.04 | .20 | −.01 | 138 | | | |
| 6. Training | .71** | .22 | .88** | .19 | −.21 | 25 | | |
| 7. Frequency | −.03 | −.27* | −.10 | −.21* | .40** | .10 | 119 | |
| 8. Judicial decision | .40** | .03 | .31* | .29** | .24** | .34† | −.02 | 295 |

[a] Phi coefficients are reported for all variables except for frequency, where point-biserial coefficients are reported. The ns for each variable appear along the diagonal.

†$p < .10$ *$p < .05$ **$p < .01$

*Tests of Hypotheses*

Table 1 presents a correlation matrix for the seven hypothesized independent variables and judicial decision. Univariate test results are shown in Table 2. Chi-square tests supported our hypotheses concerning use of job analysis (Hypothesis 1), provision of specific, written instructions (Hypothesis 3), and review of results with employees (Hypothesis 4). As expected, type of organization was not related to judicial decision. Contrary to Hypothesis 2, type of appraisal was also unrelated to judicial decision.

Concerning the second set of hypotheses, triangulation among raters was strongly related to decision, supporting Hypothesis 5. The relationship between rater training and decision (Hypothesis 6) approached significance ($p = .085$). Finally, univariate logistic regression analysis found that, contrary to Hypothesis 7, frequency of appraisal was not significantly related to decision.

*Multivariate Analysis*

Rater training had to be dropped from the LOGIT analysis due to its small sample (i.e., rater training was mentioned in only 25 cases). Significant multicollinearity was observed between job analysis and instructions ($r = .93$). In order to retain these variables in our analyses, they were combined as a two-item "scale" (alpha = .96). The remaining five variables produced a model that significantly predicted judicial decision ($\chi^2 = 34.84, p < .001$, adj. $R^2 = .15$).

The results of the LOGIT analysis can be seen in Table 3. Job analysis/instructions, employee review of results, and triangulation were all

TABLE 2

*Univariate Test Results*[a]

| Appraisal system case characteristic | $N$[b] | Number of legal cases with verdict for: Plaintiff | Defendent | $p$ |
|---|---|---|---|---|
| Job analysis (H1)? | 49 (83%) | | | .005 ($\chi^2$= 7.90) |
| No | | 16 | 5 | |
| Yes | | 10 | 18 | |
| Type of appraisal (H2)? | 167 (43%) | | | .719 ($\chi^2$= 0.13) |
| Trait | | 33 | 43 | |
| Behavior/results | | 37 | 54 | |
| Instructions given (H3)? | 50 (83%) | | | .030 ($\chi^2$= 4.71) |
| No | | 13 | 9 | |
| Yes | | 8 | 20 | |
| Employee review of results (H4)? | 156 (47%) | | | .001 ($\chi^2$= 13.47) |
| No | | 14 | 3 | |
| Yes | | 50 | 89 | |
| Triangulation among raters (H4)? | 138 (53%) | | | .005 ($\chi^2$= 7.91) |
| No | | 53 | 66 | |
| Yes | | 2 | 17 | |
| Training given (H6)? | 25 (92%) | | | .085 ($\chi^2$= 2.97) |
| No | | 12 | 6 | |
| Yes | | 2 | 5 | |
| Frequency (H7)? | 119 (60%) | | | .855 ($\chi^2$= 0.03) |
| 0 (less than once a year) | | 3 | 9 | |
| 1 (once a year) | | 27 | 49 | |
| 2 (every 6 months, up to once a year) | | 8 | 7 | |
| 3 (every 3 months, up to every 6 months) | | 2 | 9 | |
| 4 (more often than every 3 months) | | 2 | 3 | |
| Class action ? | 295 (0%) | | | .19 ($\chi^2$= 1.68) |
| No | | 99 | 150 | |
| Yes | | 23 | 23 | |
| Basis for charge ? | 294 | | | .61 ($\chi^2$= 1.81) |
| Race | (0%) | 38 | 64 | |
| Sex | | 24 | 26 | |
| Age | | 47 | 62 | |
| Other/combinations | | 13 | 20 | |

Table 2 (continued)

| Appraisal system case characteristic | $N^b$ | Number of legal cases with verdict for: Plaintiff | Defendent | $p$ |
|---|---|---|---|---|
| Purpose ? | 292 (1%) | | | .70 ($\chi^2$= 1.41) |
| Layoff | | 2 | 4 | |
| Discharge | | 65 | 96 | |
| Transfer | | 7 | 14 | |
| Promotion | | 47 | 57 | |
| Type of claim ? | 295 (0%) | | | .37 ($\chi^2$= 1.97) |
| Disparate treatment | | 112 | 161 | |
| Disparate impact | | 7 | 5 | |
| Other/combinations | | 3 | 7 | |

[a]Chi-square analyses were run for all variables except frequency (a continuous variable). Logistic regression was used for frequency.

[b]Differences among the sample sizes for each variable are due to missing data. The percentage of cases with missing data for each variable is listed in parentheses.

significant predictors of judicial decision ($p < .05$). Similar to the univariate analyses, appraisal type and frequency were not related to judicial decision, and neither were any of the control variables. An examination of the Beta-estimates indicates that defendants were more likely to win their cases when: (a) They had conducted job analysis and included written rater instructions; (b) they had allowed employees the opportunity to review appraisal results; and (c) there was evidence that more than one rater concurred with the performance assessment. A separate LOGIT analysis (not reported), where only the independent variables were used to predict judicial decision (without the controls), produced an identical pattern of results, with Beta-estimates that were virtually unchanged.

### *Discussion*

A striking feature of the present results is their consistency with Feild and Holley (1982). Despite different time periods analyzed, different court levels, and different people coding the written opinions, the two studies produced very similar results. These findings also correspond strongly to the practical recommendations found in the literature (Ashe & McRae, 1985; Barrett & Kernan, 1987; Conway, 1996; Martin et al., 1986; Veglahn, 1993). Many have lamented that keeping up with employment case law is like aiming at a moving target (Kilberg, 1988). Although changes were found for type of organization and appraisal

TABLE 3

*Results of Multivariate Logit Analysis*

| | Judicial decision[a] | |
|---|---|---|
| | *b* | *s.e.* |
| Control variables: | | |
| Year | 0.03 | .04 |
| Circuit | −0.00 | .04 |
| Class action | 0.42 | .38 |
| Basis for claim | 0.00 | .00 |
| Appraisal purpose | 0.00 | .00 |
| Type of claim | 0.63 | .49 |
| Independent variables: | | |
| Job analysis/instructions | 1.69* | .78 |
| Appraisal system type | −0.21 | .36 |
| Employee review of results | 2.04** | .70 |
| Triangulation | 2.10* | .87 |
| Frequency | −0.11 | .22 |
| −2 Log likelihood | 34.84 | |
| *df* | 11 | |
| *p*-value of chi-square statistic | .0003 | |
| adj. $R^2 = .15$ | | |

[a] Our Logit model predicts the likelihood of a judicial decision for the defendant.
*$p < .05$ **$p < .01$

type, most findings held up well over the past 15 years. This stability should increase our confidence that these are important discriminators of court outcomes in performance appraisal cases.

Specifically, both the univariate and multivariate results emphasize the importance to organizations of using job analysis, providing written instructions, and allowing employees to review appraisal results. In addition, triangulation among raters was also strongly related to judicial decision. We would note that, in a separate analysis, the number of appraisal raters was unrelated to decision (point-biserial $r = .03$, *n.s.*). Yet, in their written decisions, judges appeared to place particular emphasis on agreement among multiple raters. Further, it is interesting to note that in the multivariate analysis, employee review and evidence of triangulation were somewhat stronger predictors of judicial decision than was job analysis. The first two variables would seem to be especially salient as judges consider the procedural fairness of the appraisal systems under their review.

The findings concerning rater training approached significance in the chi-square analysis. Rater training was the only variable for which small sample size was a significant problem in the current study. In fact, an

examination of the correlation matrix in Table 1 reveals an interesting finding. That is, the first-order correlation between judicial decision and rater training, $r = .34$, although not statistically significant at $p < .05$, is second only to the correlation between judicial decision and job analysis in order of magnitude. Thus, lack of power to detect a significant relationship when one in fact existed is the most plausible explanation for our findings concerning rater training. However, it must also be noted that the $n$ we obtained for this variable is likely very close to the size of the population.[1] That is, judges do not craft their opinions so as to maximize the aims of social science research, and thus did not make frequent mention of variables such as rater training. Thus, although rater training was mentioned in fewer than 9% of the cases, when it was mentioned, almost half the time (12 out of 25) it was to note its absence, and decide for the plaintiff. Such a negative finding should alert organizations to the perils of playing "Russian roulette" with rater training.

Unlike Feild and Holley (1982), judicial decision was not significantly related to whether the appraisal system emphasized employee traits versus behaviors or results. In none of the written opinions in our sample did judges mention behavioral systems such as those recommended in the HRM literature (e.g., behaviorally anchored ratings scales, behavioral observation scales; Milkovich & Boudreau, 1994). In most cases, appraisal type was coded based upon the descriptors used to depict what was evaluated by the appraisal system. Thus, measurement imprecision may have been more of a problem for this variable than for other variables. However, another explanation is that the distinction between "traits" and "behaviors" is not especially relevant to judges. Judges generally take appraisal systems as givens, and do not wish to act as "super personnel departments" (Barrett & Kernan, 1987, p. 496). What they seek to determine is whether the system was applied fairly in the case at hand (Veglahn, 1993). Besides their simplicity of use, another reason why trait rating systems may continue to be so popular is that they can capture aspects of discretionary effort (such as citizenship behaviors), which are difficult to link directly to job descriptions, yet which are valuable for the functioning of the organization as a whole (Organ, 1988; Werner, 1994). In any case, concerns about potential inaccuracy and bias in trait-oriented systems were not reflected in our findings concerning how judges treated the type of appraisal that was conducted.

A final issue to address concerns validity. Results concerning the defendants' evidentiary burden of proof revealed that judges in this sample consistently applied the shifting burden of proof frameworks relevant to

[1] We thank an anonymous reviewer for this point.

either disparate treatment or disparate impact claims of discrimination. For example, with disparate treatment, if the organization demonstrated legitimate reasons for its actions that the plaintiff(s) could not rebut as pretextual, this led almost exclusively to decisions for defendants (and vice versa). This is desirable, because it shows that courts are following precedent, yet it is far simpler than the empirical validation efforts often called for in the literature. Several cases in this study discussed validation and statistical significance (e.g., *Kirkland v. New York State Department of Correctional Services,* 1980; *Segar v. Smith,* 1984; *Palmer v. Schultz,* 1987). Yet, all together, validation was mentioned only nine times. All nine cases were class action suits, which were decided 5/4 in favor of defendants. In these cases, validation and judicial decision were perfectly correlated. That is, in the five cases won by defendants, the defendants' validation evidence was accepted by the court. Conversely, in all four cases lost by defendants, the decisions made mention of deficiencies in (or the absence of) validation evidence. However, as argued by Barrett and Kernan (1987), it is critical to note that, in general, psychometric reliability and validity are of far less concern to judges than they are to researchers in our field. Further, the concern of Faley et al. (1984) that validity issues were not important in age discrimination cases is equally true for other types of cases.

In deciding discrimination cases, the courts placed little emphasis on validation. As Justice Brennan wrote, a plaintiff need not prove discrimination "with scientific certainty; rather, his or her burden is to prove discrimination by a preponderance of the evidence" (*Bazemore v. Friday,* 1986, p. 3009). Even more striking is Justice O'Connor's statement in *Watson* (1988, p. 2790) that "employers are not required, even when defending standardized or objective tests, to introduce formal 'validation studies' showing that particular criteria predict actual on-the-job performance." This statement is contrary to the *Uniform Guidelines.* It is not, however, contrary to court practices of paying more attention to the "fairness" than to the validity of employment practices (Lee, 1989). Although some have argued that the *Watson* decision showed the Supreme Court moved away from the "great deference" it once accorded EEOC guidelines (Kandel, 1988), this study suggests that the high court simply endorsed what the lower courts have practiced all along. In *Ward's Cove Packing v. Atonio* (1989, pp. 2125–2126), Justice White wrote that "the dispositive issue is whether a challenged practice serves, in a significant way, the legitimate employment goals of the employer.... The touchstone of this inquiry is a reasoned review of the employer's justification for his use of the challenged practice."

*Caveats, Limitations, and Future Directions*

Several points are in order to put our findings in context. First, our primary goal was to determine which variables explained courts of appeals decisions. We emphasize that our study does not address what is "legal" versus "illegal" concerning performance appraisal. Our findings suggest that organizations are more likely to experience favorable outcomes in legal proceedings if certain characteristics are present in their appraisal systems. However, if they fail to meet a *matter of law,* such as meeting the appropriate burden of proof, then no amount of these appraisal system characteristics can overcome such a deficiency (Roehling, 1993). On a related note, different legal issues arise in various types of cases, such as disparate treatment versus disparate impact claims, or when appraisals are used for different purposes (Martin & Bartol, 1991). As a whole, our control variables were unrelated to decision pattern, which increases the confidence one can have in our statistical analyses. However, we in no way wish to minimize the different legal procedures that apply to different types of appraisal court cases. Readers are advised to seek legal counsel, and to consult other sources for guidance with specific cases (Ashe & McRae, 1985; Ledvinka & Scarpello, 1991; Martin & Bartol, 1991; Roehling, 1993).

Further, we are emphatically not arguing that variables are unimportant if no statistically significant relationship was observed between that variable and judicial decision. Obviously, organizations are well advised to provide frequent appraisals and to train their raters, regardless of whether statistical tests were significant or not. However, just as meta-analytic findings can supplement narrative literature reviews, statistical analyses can add to the information provided by narrative reviews of appraisal court cases. In particular, it is important to see how often judges do note factors such as agreement among multiple raters, employee review of results, and even job analysis and written instructions. In those (albeit rare) cases in which organizations face court proceedings where performance appraisal is at issue, our results are instructive concerning which variables are most likely to be noted in court opinions. Although it is unclear the extent to which our findings would generalize to decisions decided in state courts (Roehling, 1993), carryover to federal district court decisions is likely (cf. Feild & Holley, 1982).

As in Feild and Holley (1982), the biggest problem using this methodology was the large amount of missing data. Even with our larger sample of court cases, this remained an issue in the current study as well. Caution is warranted concerning our chi-square analyses, because in several instances, at least one of the cell frequencies fell below the recommended minimum of five cases per cell. For our multivariate analyses,

the small sample sizes for some independent variables, coupled with our use of mean substitution could also raise concerns. However, we would note that our multivariate results parallel the univariate results quite closely, which suggests that our multivariate results were not distorted by our use of mean substitution for missing data.

Future research should study both federal district court and state court cases. Second, this study did not include any cases concerning university tenure decisions (e.g., *University of Pennsylvania v. EEOC,* 1990). As the number of such cases builds, this should be a profitable area of future study as well. Third, quasi-field studies are strongly recommended, (e.g., where lawyers serve as subjects who make "judgments" based on case summaries; Roehling, 1993).

In closing, we would stress that accuracy and fairness should not be viewed as an either/or proposition. Both are necessary and important. This argument has been made before (Folger et al., 1992), but takes on added significance when the context of study is performance appraisal court cases. Issues of justice and due process have long been central to the literature on unions and labor relations (Fossum, 1995; Wheeler & Rojot, 1992). Further, grievance procedures have been advocated in non-union settings as well, to foster "corporate due process" (Ewing, 1989). With this history, the relatively recent "revelation" concerning the importance of due process in performance appraisal is itself revealing. As Veglahn (1993 p. 600) noted in his review of performance appraisal court decisions, "the issue of procedural fairness of the system is examined by the courts much more closely than the issue of accuracy of performance evaluation." We do not think that due process considerations can or should supplant concerns for accuracy; issues of accuracy and fairness should rather supplement one another (Folger et al., 1992). By using both accuracy and due process perspectives, we believe this study makes an important contribution toward understanding the issues involved when performance appraisal is relevant to an employment discrimination judicial opinion.

## REFERENCES

Amicus curiae brief for the American Psychological Association; In the Supreme Court of the United States: Clara Watson v. Fort Worth Bank & Trust. (1988). *American Psychologist, 43,* 1019–1028.

Ashe RL, Jr., McRae GS. (1985). Performance evaluations go to court in the 1980s. *Mercer Law Review, 36,* 887–905.

Barrett GV, Kernan MC. (1987). Performance appraisal and termination: A review of court decisions since Brito v. Zia with implications for personnel practices. PERSONNEL PSYCHOLOGY, *40,* 489–503.

Bazemore v. Friday, 106 S. Ct. 3000 (1986).

Bazerman MH. (1990). *Judgment in managerial decision making.* NewYork: John Wiley & Sons.

Bernardin HJ, Beatty RW, Jensen W. (1980). The new uniform guidelines on employee selection procedures in the context of university personnel decisions. PERSONNEL PSYCHOLOGY, *33,* 301–316.

Brito v. Zia Company, 478 F.2d 1200 (1973).

Cardy RL, Keefe TJ. (1994). Observational purpose and evaluative articulation in frame-of-reference training: The effects of alternative processing modes on rating accuracy. *Organizational Behavior and Human Decision Processes, 57,* 338–357.

Cascio WF, Bernardin HJ. (1981). Implications of performance appraisal litigation for personnel decisions. PERSONNEL PSYCHOLOGY, *34,* 211–226.

Cohen J, Cohen P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences,* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen ML, Berring RC. (1983). *How to find the law,* (8th ed.). St. Paul, MN: West.

Conway JM. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22,* 139–162.

Donner A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician, 36,* 378–381.

Ewing DW. (1989). *Justice on the job: Resolving grievances in the nonunion workplace.* Boston: Harvard Business School Press.

Faley RH, Kleiman LS, Lengnickell ML. (1984). Age discrimination and personnel psychology: A review and synthesis of the legal literature with implications for future research. PERSONNEL PSYCHOLOGY, *37,* 327–350.

Feild HS, Holley WH. (1982). The relationship of performance appraisal system characteristics to verdicts in selected employment discrimination cases. *Academy of Management Journal, 25,* 392–406.

Feild HS, Thompson DT. (1984). Study of court decisions in cases involving employee performance appraisal systems. *Daily Labor Report* (BNA), 248, E-1, 12/26.

Folger R, Konovsky MA. (1989). Effects of procedural and distributive justice on reactions to pay raise decisions. *Academy of Management Journal, 32,* 115–130.

Folger R, Konovsky MA, Cropanzano R. (1992). A due process metaphor for performance appraisal. In Staw B, Cummings L (Eds.), *Research in Organizational Behavior* (Vol. 14, pp. 129–177). Greenwich, CT: JAI Press.

Forkosch MD. (1958). American democracy and procedural due process. *Brooklyn Law Review, 24,* 173–253.

Fossum JA. (1995). *Labor relations: Development, structure, process* (6th ed.). Homewood, IL: Irwin.

Goldman S. (1975). Voting behavior on the United States Courts of Appeals revisited. *American Political Science Review, 69,* 491–506.

Ilgen DR, Barnes-Farrell JL, McKellen DB. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321–368.

Kandel WL. (1988). Current developments in employment litigation. *Employee Relations Law Journal, 14,* 263–275.

Kilberg WL. (1988). From the editor. *Employee Relations Law Journal, 14,* 159–161.

Kirkland v. New York State Department of Correctional Services, 628 F.2d 796 (1980).

Korsgaard MA, Roberson L, Rymph D. (1996, April). *Promoting fairness through subordinate training: The impact of communication style on manager's effectiveness.* Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, Inc., San Diego.

Landy FJ, Barnes JL, Murphy KR. (1978). Correlates of perceived fairness and accuracy of performance evaluation. *Journal of Applied Psychology, 63,* 751–754.

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

Ledvinka J, Scarpello VG. (1991). *Federal regulation of personnel and human resource management* (2nd ed.). Boston: PWS-Kent.

Lee BA. (1989). Subjective employment practices and disparate impact: Unresolved issues. *Employee Relations Law Journal, 15,* 403–417.

Lind EA, Tyler TR. (1988). *The social psychology of procedural justice*. New York: Plenum.

Longenecker CO, Gioia DA, Sims HP. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1,* 183–193.

Martin DC, Bartol KM. (1991). The legal ramifications of performance appraisal: An update. *Employee Relations Law Journal, 17,* 257–286.

Martin DC, Bartol KM, Levine MJ. (1986). The legal ramifications of performance appraisal. *Employee Relations Law Journal, 12,* 370–396.

McDonnell-Douglas Corp. v. Green, 93 S. Ct. 1817 (1973).

Milkovich GT, Boudreau JW. (1994). *Human resource management,* (7th ed.). Burr Ridge, IL: Irwin.

Miller CS, Kaspin JA, Schuster MH. (1990). The impact of performance appraisal methods of age discrimination in employment act cases. PERSONNEL PSYCHOLOGY, *43,* 555–578.

Nagel SS, Neef MG. (1979). *Decision theory and the legal process.* Lexington, Mass.: D. C. Heath.

Organ DW. (1988). *Organizational citizenship behavior: The good soldier syndrome.* Lexington, MA: Lexington Books.

Palmer v. Schultz, 815 F. 2d 84 (1984).

Pulakos ED. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581–588.

Pulakos ED. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38,* 76–91.

Reynolds WL, Richman WM. (1981). An evaluation of limited publication in the United States Courts of Appeals: The price of reform. *U. of Chicago Law Review, 48,* 573.

Roehling MV. (1993). "Extracting" policy from judicial opinions: The dangers of policy capturing in a field setting. PERSONNEL PSYCHOLOGY, *46,* 477–502.

Schneier DB. (1978). The impact of EEO legislation on performance appraisals. *Personnel, 55,* 24–34.

Schuster MH, Miller CS. (1981). Performance evaluations as evidence in ADEA cases. *Employee Relations Law Journal, 6,* 561–583.

Schuster M, Miller CS. (1984). An empirical assessment of the Age Discrimination in Employment Act. *Industrial and Labor Relations Review, 38,* 64–74.

Segar v. Smith, 738 F.2d 1249 (1984).

Spaeth HJ. (1985). Supreme court disposition of federal circuit court decisions. *Judicature, 68,* 245–250.

Sulsky LM, Day DV. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79,* 535–543.

Taylor MS, Tracy KB, Renard MK, Harrison JK, Carroll SJ. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly, 40,* 495–523.

Texas Department of Community Affairs v. Burdine, 101 S. Ct. 1089 (1981).

Thibaut J, Walker L. (1978). A theory of procedure. *California Law Review, 66,* 541–566.

Tsui AS. (1990). A multiple-constituency model of effectiveness: An empirical examination at the human resource subunit level. *Administrative Science Quarterly, 35,* 458–483.

Uniform guidelines on employee selection procedures (1978). *Federal Register, 43,* 38290–38315.

University of Pennsylvania v. EEOC, 110 S. Ct. 577 (1990).

Veglahn PA. (1993). Key issues in performance appraisal challenges: Evidence from court and arbitration decisions. *Labor Law Journal, October,* 595–606.

Wards Cove Packing Company v. Atonio, 109 S. Ct. 2115 (1989).

Watson v. Fort Worth Bank and Trust, 108 S. Ct. 2777 (1988).

Werner JM. (1994). Dimensions that make a difference: Examining the impact of in-role and extrarole behaviors on supervisory ratings. *Journal of Applied Psychology, 79,* 98–107.

Wheeler HN, Rojot J, (Eds.). (1992). *Workplace justice: Employment obligations in international perspective.* Columbia, SC: University of South Carolina Press.