A Behavior-based Evaluation Instrument for Judges

E. Lee Bernick and David J. Pratto

The belief that judicial performance should be evaluated has gained increasing momentum. A number of states have used surveys of attorneys as a primary source of information about judicial performance. The evaluation survey, if not carefully constructed, may lead to biased evaluations. This research reports on an effort to create a survey that would be easily administered while at the same time providing information free from bias. Using a procedure developed in other occupational fields, the researchers developed an instrument that uses "critical incidents" of actual judicial behavior as benchmarks for scales to measure judicial performance across six dimensions. Each of the six dimensions consists of five items. The instrument was field tested and shown to be free from bias (i.e., Halo/Horn and leniency effects) often found in evaluation instruments.

The people of the United States believe that voting for public officials is the ultimate expression of democracy. In some states, this adherence to the democratic ideal means judges are elected rather than appointed to office (these elections may be partisan or nonpartisan). Other states initially appoint judges and still provide for a sense of democratic accountability by holding retention elections. Thus, on some periodic basis, voters decide whether to retain a judge.

Regardless of the circumstances, we expect the electorate to become informed about the candidates and the issues. However, research has found that the citizenry pays little heed to judicial elections (Jacob, 1990). Complicating the issue for the voters is the lack of information available regarding judicial candidates. Underlying this problem in many cases is the strong norm to protect the judicial system's independence, including a desire to permit judges to make decisions without fear of retribution nor favoritism (Farthing-Capowich, 1985). In many states, rules of judicial conduct severely limit what can be said by or about a jurist or his or her judicial record. As a result, there has been a strong desire
to create a system that will provide information to the voter while maintaining the
independence of the courts.

One mechanism touted as able to meet the balancing test between judicial indepen­
dence and an informed electorate has been some form of systematic judicial performance
review (Keilitz and McBride, 1992). Evaluation of judicial performance came of age in the
1980s, much of it through the work conducted by the Evaluation of Judicial Performance
Project (Farthing-Capowich and McBride, 1987). By 1992, fourteen states had some
formal program of judicial evaluation. While there has been considerable growth in the
development of these programs, they are not without problems.

Many of the problems focus on the implementation of the evaluation system. Ques­
tions have been raised: 1) Who will evaluate the jurist? 2) Who will sponsor the evaluation
program? 3) What issues or facets of the job should be evaluated? and 4) What evaluation
instrument will be used? (Farthing-Capowich, 1984). These questions provide an outline
of the issues germane to the validity of the process. The last two questions are the foci of
this article.

The validity of any process or measure may be biased by many factors. Bias in
performance evaluation can occur from the procedures used to conduct the evaluation (e.g.
sampling bias), from individual responses, or from the evaluation instrument itself (Groves,
1989). Thus, one of the essential questions in the construction of a judicial evaluation
program is: “Can we develop a performance evaluation instrument free from bias?” This
article reports on research that provides information on an evaluation instrument created to
overcome shortcomings inherent in many evaluation instruments. The instrument uses
“critical incidents” of actual judicial behavior that tap six dimensions of judicial perfor­
mance. Results of an application of the instrument to test for response bias as well as for
how well the instrument discriminates between good and bad judicial behavior are reported.

Current Evaluation Systems

There is no single method to collect information about judicial performance (Keilitz
and McBride, 1992). Some systems analyze court records for the outcome of cases before
a judge, other systems ask for information directly from judges, while still others employ
retired justices to observe a sitting judge’s court. A favorite technique is to use surveys,
including surveys of peace officers, jurors, litigants, and lawyers (Farthing-Capowich,
1984; Keilitz and McBride, 1992). While these populations are perceived to be the most
knowledgeable, some individuals express concern about their ability to evaluate judges
objectively. Thus, an important consideration in the creation of any evaluation procedure
is the development of an instrument appropriate for the population. This research focuses
on the development of a survey instrument for attorneys. Attorneys are used most
frequently in evaluating judges because the record is clear that they will be cooperative and
act responsibly in their efforts to evaluate judges (Handler, 1986).

The evaluation surveys ask attorneys a series of questions that represent dimensions
(i.e., areas of performance) essential for the job of “Judge.” The number of dimensions
evaluated vary; they ranged from as few as three in New Jersey to as many as fourteen in
Alaska (Keilitz and McBride, 1992). The New Jersey system, when it was first initiated, tried to measure "comportment," "judicial management skills," and "legal ability of judges" (Handler, 1984: 41). In Vermont judges have been evaluated on ten dimensions, including a judge’s involvement in the community and media relations (Farthing-Capowich, 1984). Alaska’s system has sought to evaluate even more dimensions, including settlement skills, courtesy, and human understanding and compassion (Bremson, 1984). Thus, no single set of items has been adopted for use across all the states that have evaluation systems.

In a somewhat similar fashion, the method employed to actually measure attorneys’ assessments on these dimensions vary. Nevertheless, most evaluations employ a set of "fixed choice" items. In these situations, attorneys respond to statements about judicial behavior by rating the judge on a scale. Evaluation instruments have used scales with as few as three options and as many as nine. Alaska is somewhat typical in its use of a five-point scale in which 1 = "unacceptable" and 5 = "excellent" (Bremson, 1984).

Research on performance review instruments indicates that these scaling procedures are not without the potential for bias. One potential source of bias rests with the requirement that each assessor must define the adjective; in other words, what does "unacceptable" mean. Without a specific referent point, two individuals may observe the same behavior and evaluate it differently. Moreover, we know from other research on performance appraisals that a variety of biases may creep into the process (Cascio, 1978). Some evaluators have a tendency to rate everyone the same; thus, they are not discriminating in their assessment. Among some assessors this results in a tendency to evaluate everyone in the middle—the "central tendency" bias (Latham and Wexley, 1981). Other appraisers give everyone high marks (the "leniency" error), while still others consistently evaluate too harshly, resulting in low scores for all appraisees (Latham and Wexley, 1981).

Another source of bias among appraisers is the "Halo Effect" (Blanz and Ghiselli, 1972). In the "Halo Effect," individuals transfer the positive assessment on the first dimension of judicial behavior to all of the other dimensions. An evaluator's initial assessment in one area results in high marks across the board for a judge. Thus, if "knowledge of the law" was the first dimension and a judge was excellent in this area, an attorney may assume the judge is also excellent in other areas; for example, "courtesy." The converse of the "Halo Effect" is known as the "Horns Effect" (a poor initial assessment on one dimension leads to all negative assessments). In a somewhat similar vein, lawyers may err in attempting to make "logical connections"; if they rate a judge high on one dimension, then they assume that a judge performs a logically related task at an equally high level (Cascio, 1978: 322). For example, a positive evaluation on "judicial favoritism" may lead to a positive evaluation on a seemingly related item: "prejudging."

To overcome these biases, industrial psychologists in other fields have developed behavior-anchored rating scales (BARS) or behavior-expected scales (BES) (Latham and Wexley, 1981). These scales are also designed to improve interrater reliability. The essential component of this form of evaluation instrument is the use of actual behavioral examples to anchor or define the points on a rating scale. "Behavioral descriptors, exemplifying various degrees of each dimension, are printed beside the line at different heights according to a scale position" (Smith and Kendall, 1963: 150). Thus, instead of the
adjective "unacceptable," a specific example of unacceptable behavior would be placed next to the scale score of "1."

Scales using behaviors as anchors are an adaptation of an assessment process known as "critical incidents" (Campbell et al., 1973). The scales are created by using examples (i.e., "critical incidents") of behavior exhibited by people doing the job, in this case a judge. As such, the scales are designed with a set of common experiences. In constructing the scales, extensive participation on the part of people who are involved in the work enables the scales to have the actual language of the occupation and not some generic nomenclature (Campbell et al., 1973). All of this means that behavior-anchored rating scales have high content validity (Smith and Kendall, 1963). (Some individuals correctly point out that it is more accurate to refer to these scales as behavior-expected scales—"the examples . . . represent not actual observed behavior but inferences or predictions [expectations] from observations" [Smith and Kendall, 1963: 150].) Behavior-based scales not only have high content validity, but they also have been shown to reduce the Halo/Horn Effect, reduce leniency error, and encourage appraisers to make distinctions in behavior (Campbell et al., 1973).

In sum, there is an approach to designing an appraisal instrument for judges that increases the validity of the appraisal process. Rather than requiring appraisers to use abstract adjectives in their assessment of judges, scales can be developed that permit appraisers to assess judges by using examples of actual behavior. This procedure can provide more, and more meaningful, information about judicial behavior. Moreover, the process of developing the scales will ensure that the judicial behavior cited is deemed important to evaluate.

Creation of the Instrument

A core panel of nine attorneys, two former judges, and three professional researchers was created to develop the evaluation instrument. The attorneys on the panel came from two different cities within a large urban county. They were selected by the sponsoring organization because of their well-established reputations among corporate and public sector attorneys and among attorneys in private practice. The panel served in an advisory capacity as well as in the direct preparation of the instrument. The panel’s first task was to develop a list of 125 attorneys, who they knew had courtroom experience. Moreover, the panel was instructed to select attorneys that would ensure representation from all facets of the law and from diverse social groups.

These 125 attorneys were contacted by mail and asked to assist in the development of an evaluation instrument. They were directed to think about their experiences with judges and requested to provide examples of both good and bad behavior on the part of judges that they had observed. Each individual was sent two sets of five color-coded index cards. The attorneys were asked to give no more than five examples (although they did not have to give five) of good behavior and five examples of bad behavior. The attorneys were also instructed that these examples should not be limited to just one specific area of a judge’s job.

More than 300 examples of behavior were returned from 34 attorneys (who were determined to be representative of the original 125 contacted). The researchers reviewed the responses and eliminated any two cards that contained identical wording.
remained were 238 examples of judicial behavior, including 109 positive examples and 129 negative examples. These 238 statements were reviewed by an executive committee (consisting of the researchers and two of the advisory panel’s attorneys) to eliminate examples that were not relevant to judicial proceedings, essentially the same in meaning, or had little generalizability (the examples were too specific to a particular case). Through this process, the number of statements was reduced to 101. These 101 statements then needed to be arranged into dimensions.

The executive committee, using information from other evaluation instruments, culled out seven potential dimensions: 1) Courtesy, 2) Application of the Law, 3) Impartiality, 4) Case Load Management, 5) Diligence (Hard Work), 6) Respect for the Judicial System, and 7) Judicial Temperament. The seven dimensions were selected because of their commonality of citation in the evaluation instruments reviewed. The next task was to determine the relationship between the seven dimensions and the 101 behavioral statements.

The core panel members were mailed the 101 statements with instructions to read and place each statement into one of the seven dimensions. They could place a statement into only one dimension; however, if they felt a statement did not represent any single dimension, they were advised to place an “X” next to the statement. The panel members also had the opportunity to indicate that a statement embodied some other facet of judicial behavior (none were, in fact, identified).

The next step was to analyze the results concerning the proper placement of each statement into one of the dimensions. Statements were eliminated if no agreement existed regarding placement into a dimension. Specifically, if less than 75 percent of the respondents agreed on the placement of a statement into a particular dimension, it was considered a poor example of a particular kind of judicial behavior. Moreover, the adequacy or appropriateness of a dimension was questioned if a marginal number of statements were placed into the proposed dimension. At the completion of this stage of the analysis, 52 of the 101 statements were placed into one of the seven dimensions. Five of the seven dimensions appeared to be useful categories containing at least five statements each. The other two dimensions (Case Load Management and Diligence) had fewer than five items each. Rather than eliminating them outright, it was determined that the core panel would review these dimensions in the next phase of the development.

In the next phase, the core panel ranked each statement within a dimension, with the best example of judicial behavior ranked first and the worst example ranked last. The ranking process resulted in the elimination of unnecessary or inappropriate items and the development of a five-item scale. An item was eliminated if there was little agreement among the panelists on its specific rank. The mean score and the standard deviation for each item was used to place the items on the scale; the items were ranked based on their mean scores. However, statements with large standard deviations were eliminated. This process provided five items for five of the scales.

Review of the two potential dimensions with less than five examples resulted in the elimination of Case Load Management because the panel did not believe it could be made into a credible dimension. The panel determined that the proposed Diligence dimension was important and could be made into a scale with one additional item. Accordingly, the panel developed a fifth item to be used to complete the full five-item dimension. This
process yielded six dimensions with scales consisting of five examples of behavior in each that represented different levels of performance.

After the completion of the six dimensions, the advisory panel reviewed a final draft of the evaluation instrument. The executive committee met and made final corrections and additions to the instrument, which was then ready for field-testing in a large urban county in North Carolina.

Two modifications to the traditional behavior-anchored rating scales were made in the development of this evaluation instrument. First, to reduce the potential for appraiser error, no numbers were placed next to any of the statements in a dimension. By not placing numbers next to the statements individual appraisers could not assume that there was an order of high to low to the statements. Thus, there should be a reduction of error due to leniency or harshness. Second, to further reduce the potential for error, the statements were not placed within the dimension in any order. Thus, the statement representing the worst behavior could be the middle choice of the five provided. On another dimension, the statement representing the best behavior may have been in the middle. In Table 1, we provide the five statements for each of the six dimensions (the number in parentheses provides the actual order of the statements).

Testing the Evaluation Instrument

To test the instrument, it was decided that a mail survey of attorneys would be used. The survey’s instrument consisted of a single page for each judge to be evaluated. All district court judges (ten) and all superior court judges (twelve) in one county of North Carolina were evaluated.

The population of appraisers consisted of all attorneys (998) in one large urban county in North Carolina. Each attorney was asked to evaluate every district court judge and six randomly assigned superior court judges. Each attorney was asked to complete the evaluation only for those judges that he or she had direct experience in the past two years. Attorneys who had not appeared before a particular judge in the past year were instructed to exclude themselves from evaluating that judge. Additional research, in conjunction with judicial evaluation surveys, has shown that only 56 percent of all attorneys who are members of the bar associations appear in court on a regular basis (Bernick and Pratto, 1994). Therefore, a more accurate number of potential respondents (attorneys) would be 559 and not the 998 in our original sample.

The survey process included three mailings. At the completion of the three mailings, 482 respondents out of the possible 998 attorneys had responded—a response rate of 48 percent. As expected, not all of the respondents could have evaluated the judges (196...
Table 1  
Judicial Dimension Scales*  

APPLICATION OF THE LAW: The Judge . . .  
follows legal precedent even when it goes against the judge's strongly held personal belief. (1)  
knowingly refuses to follow the rules or makes rulings that are contrary to law. (5)  
ignores the law and rules according to outside nonevidentiary factors. (4)  
usually matches appropriate law to facts. (3)  
makes every effort to make the correct decision based upon the law. (2)  

COURTESY: The Judge . . .  
does not show respect for parties or witnesses. (5)  
is inconsiderate of jury's time and service. (3)  
willingly grants recesses to allow witnesses to "regain their composure." (2)  
works through lunch breaks and does not consider impact on court employees and people in the courtroom. (4)  
has a pleasant manner toward people. (1)  

IMPARTIALITY: The Judge . . .  
shows no favoritism to either side in a lawsuit. (1)  
treats parties differently according to their appearance, lifestyle, or personal views. (3)  
does not allow judge's personal background to influence rulings. (2)  
decides favorably for lawyers because of their political support. (5)  
reveals personal bias toward one side in a case. (4)  

JUDICIAL TEMPERAMENT: The Judge . . .  
listens patiently to all testimony and arguments before issuing a ruling. (1)  
uses intimidation to maintain control of the courtroom. (4)  
appears most of the time to be bored with a case. (3)  
always maintains a professional demeanor in the courtroom. (2)  
frequently loses temper or retaliates through decisions. (5)  

RESPECT FOR JUDICIAL SYSTEM: The Judge . . .  
does not interject personal comments in the courtroom. (1)  
uses judicial discretion to reach a fair decision. (3)  
ensures that an unrepresented party's rights are protected. (2)  
scares parties into settling a case. (4)  
is influenced by ex-parte activities. (5)  

DILIGENCE: The Judge . . .  
stays abreast of the law. (3)  
puts personal schedule ahead of court schedule. (5)  
shows a willingness to hear complicated and time-consuming cases. (2)  
reads files and materials submitted by counsel. (1)  
does not use court time effectively. (4)  

* The items are presented as they were in the actual evaluation form. The numbers, not printed on the evaluation instrument, are added to indicate item placement 1 (best) to 5 (worst) on each dimensional scale. Copyright, Court Watch of North Carolina, Inc.
attorneys indicated that they had not appeared in court in the past two years). Thus, the total number of attorneys who actually evaluated justices was 286, which represents 51 percent of the predicted likely respondents (559).

Results of Field-testing

After receiving the completed evaluation forms, the analysis turned to assessing the psychometric properties of the evaluation instrument. First, the instrument was evaluated to determine if there was a tendency for attorneys to be too lenient—to give overly favorable ratings. Typically, leniency would be considered present if the mean score for judges was not at the midpoint of a scale and the deviation was toward a more favorable rating. This definition of leniency assumes the group being evaluated resembles a normal population. However, given the nature of the population (i.e., judges), one would not expect "average behavior!" In fact, the curve for judges can be expected to be skewed toward the more favorable side of the rating scale.

As noted previously, the six dimensions were made up of five items each. The scale scores could range from 1.00 to 5.00, with 1.00 being the most favorable scale score. The mean scores range from 1.65 to 2.68 with three below the scale score of 2.00 and three above 2.00 (see Table 2). Thus, there is a tendency for generally favorable ratings, but the mean scale scores do not appear to be excessively lenient. The mean scores indicate a willingness on the part of attorneys to evaluate some judges as less than "perfect."

The standard deviations for the six dimensions are all less than 1 and could be misconstrued as supporting a tendency for raters to be lenient. As noted previously, the range on the scale ran from 1.00 to 3.00 rather than the theoretical 1.00 to 5.00. While we cannot say for certain, the evidence from this initial survey would support the theory that the scales were more likely to have a much smaller range than the potential 1.00 to 5.00. We recognize the tentative nature of this judgment, but would suggest that the narrower range reflects the uniqueness of judges as a population. Thus, the potential distribution will have a smaller standard deviation and yet can still discriminate among judges.

In fact, that is what did occur, and the six dimensions do appear to discriminate among judges. Analysis of nineteen of the twenty-two judges across the six dimensions reveals the presence of at least one judge in each dimension to be one standard deviation beyond the mean. Specifically, in 11 percent of the evaluations, judges were placed one standard deviation beyond the mean, and in each instance the direction was toward the negative end of the scale. In sum, the information supports the conclusion that no leniency effect exists in this instrument and that attorneys do discriminate in their assessment of judges (see Table 2).

Our definition of the Halo/Horn Effect maintains that this form of bias exists if the attorney uses the evaluation given on the first dimension to determine all of the remaining

---

3 For some elements of the analysis, we eliminated three of the judges because the number of attorneys evaluating the judge was too small for meaningful analysis.
Table 2
Means and Standard Deviations for Each Judicial Performance Dimension

<table>
<thead>
<tr>
<th>Performance Dimension</th>
<th>Mean</th>
<th>Standard Deviation (SD)</th>
<th>Number of Judges Beyond 1 SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>2.50</td>
<td>.568</td>
<td>2</td>
</tr>
<tr>
<td>Impartiality</td>
<td>1.77</td>
<td>.686</td>
<td>2</td>
</tr>
<tr>
<td>Respect for System</td>
<td>2.68</td>
<td>.671</td>
<td>2</td>
</tr>
<tr>
<td>Courtesy</td>
<td>1.65</td>
<td>.692</td>
<td>3</td>
</tr>
<tr>
<td>Judicial Temperament</td>
<td>1.79</td>
<td>.615</td>
<td>3</td>
</tr>
<tr>
<td>Diligence</td>
<td>2.65</td>
<td>.710</td>
<td>1</td>
</tr>
</tbody>
</table>

scores. Thus, the attorney places a judge at the same level for all six dimensions. To determine the existence of this form of bias, we screened all the separate 2,178 evaluations of all the judges. We examined the pattern of responses for the six dimensions for each of the 2,178 evaluations and found 26 evaluations where attorneys gave uniformly positive scores (1s) across all six dimensions. In three instances, the attorneys gave uniformly negative scores (5s) across the six dimensions. Thus, only 29 of the 2,178 evaluations (or 1.3 percent) exhibited the Halo/Horn Effect. Clearly, the Halo/Horn Effect is insignificant in the evaluation instrument. The analysis lends further evidence of the power of this instrument to encourage attorneys to discriminate in their assessment of judicial behavior.

Our final assessment of the instrument looked at interrater reliability. We measured the degree of agreement between attorneys on their assessment of judicial performance. Lack of attorney agreement would call into question the ability of the dimension to accurately reflect the judicial behavior being measured.

We examined the level of agreement for the attorneys of each judge for each dimension. Thus, we looked at the population evaluating each of the nineteen judges across the six dimensions and divided the population of attorneys for each judge into two random groups. We then tested for the difference of means between the two groups across each dimension using a t-test. There were 114 comparisons, and in only six instances were the differences between the two groups significant at the .05 level. Thus, in 95 percent of the comparisons no differences between the two groups of attorneys were found. These results provide support for the robustness of this judicial evaluation instrument in its sensitivity across dimensions of judicial behavior and its ability to evaluate the performance of judges.

Conclusions

The evaluation of judicial performance has become more prevalent across the states. One typical way to evaluate judicial performance is to use survey instruments that resemble a performance appraisal instrument. In many of the states, these surveys are sent to attorneys. One of the major drawbacks of survey instruments, as they are generally developed, is the potential for bias.
This research is a response to the need to develop an evaluation instrument relatively free from the bias that typically occurs in performance appraisal instruments. Using a procedure developed in other occupational fields, the researchers developed an instrument that uses "critical incidents" of actual judicial behavior as benchmarks for scales to measure judicial performance across six dimensions. The final instrument asked attorneys to evaluate judges on six dimensions with a scale that had five examples of judicial performance; the incidents used in the scale for each dimension represented a continuum that ran from very good to very poor behavior.

The instrument was field tested in a large urban county for twenty-two judges. The results indicate that the instrument discriminates in judicial behavior and does not manifest the Halo/Horn Effect or leniency bias. Moreover, the analysis provided evidence of the reliability of the assessment across attorneys. In sum, attorneys did not give uniformly high or low marks, were consistent in their assessment of judges, and yet were able to discriminate between good and bad judicial behavior.

Another test for any evaluation instrument is the willingness of others to accept its validity. Moreover, the results of the instrument should be used to improve the judicial system. First, many attorneys in the county provided feedback that the rank order of the judges corresponded to their general understanding of the judges' performance. In other words, the judges who ranked first and second were generally considered the best while the judges ranked at the bottom were considered inferior. Second, the local newspapers gave widespread coverage of the results and used the findings to make recommendations in their editorials endorsing judicial candidates. Third, although we cannot say this with certainty, the public may have accepted the evaluations. A possible consequence of this evaluation and its coverage was the electoral defeat of the judge rated lowest. (The defeated judge did indicate that the evaluation was one possible explanation for his defeat.) Finally, we have evidence that the judges gave serious consideration to the rankings and the instrument. One judge, who eventually moved to the state court of appeals, wrote a letter rationalizing his position in the rankings and offered an interpretation of the scales. Two other judges asked to meet with the research team to gain a better understanding of the process and how to interpret the results. Moreover, they offered suggestions to the research team for future evaluations. We believe the evaluation procedure and the scales are acceptable to diverse groups in the community.

In sum, this new performance instrument has several important features that warrant its use. First, it allows for the evaluation of judicial performance using real behavior rather than the vague and ill-defined scaling adjectives that some instruments use. Second, it is free from the typical biases that are found in many performance evaluation instruments. Third, as with some other instruments, it is easily administered, encouraging high levels of response. As a result of the previously cited benefits, the resulting evaluation instrument provides substantial amounts of information to both the individuals being evaluated and the individuals making decisions about the retention of judges, whether they are legislators, other reviewing groups, or the voters themselves.

Finally, if a group wants to create their own behavior-based scales instead of using the ones presented, we encourage them to follow the procedures outline in this article. The authors would be more than willing to assist with technical advice. 

182
REFERENCES


