Improving Judicial-Performance Evaluation:

Countering Bias and Exploring New Methods

Jennifer K. Elek & David B. Rottman

Official judicial-performance evaluation (JPE) programs in the United States emerged to achieve important judicial-branch objectives. JPE programs respond to the need for courts to demonstrate accountability, provide information for voters in low-information judicial-retention elections, improve the quality of the bench by providing feedback for individual judges to use for self-evaluation purposes, and assist judicial administrators in making decisions on retention and assignments in some states with appointed judiciaries. A number of professional organizations, such as the American Bar Association, American Judicature Society, and the Institute for the Advancement of the American Legal System, are strong advocates for the value of JPE programs.[1]

Eighteen states and the District of Columbia currently operate official JPE programs, mostly conducted by the judicial branch itself, but some conducted by executive-branch agencies in a few states.[2] Whether official or unofficial, nearly all JPE programs rely upon surveys distributed to attorneys and court staff—and in some instances to jurors, litigants, and others—as the exclusive or a primary method for measuring judicial performance.[3] Most state JPE programs are based on the American Bar Association's Black Letter Guidelines for the Evaluation of Judicial Performance,[4] and several states use some variation of the model surveys put forth by the ABA Lawyers' Conference.[5]

The great potential of JPE programs to improve the quality of justice is not being realized due to fundamental problems in the evaluation methodologies they use. Some of these problems result from deficiencies in basic survey design or in the manner in which the surveys are distributed, issues we have previously addressed.[6] Other problems result from the failure to incorporate efforts to minimize the potential for systematic biases against women and minority judges in JPE survey ratings.

The good news is that if states adopt best practices in survey design generally and work-performance surveys in particular, it is possible to improve the validity of JPE surveys and minimize the presence of bias in evaluation ratings. This article briefly explains the potential for bias in JPE surveys and then describes one effort to design a new JPE survey that achieves the above goals.

CONCERNS OF BIAS IN SURVEY-BASED JUDICIAL-PERFORMANCE EVALUATION

Anecdotally, concerns about the problem of gender and racial bias in results of state JPE surveys have been voiced for decades.[7] Unfortunately, very little research has been conducted on the efficacy of state JPE survey instruments.[8] It was not until 2011, when researchers at the University of Nevada published evidence of systematic gender and racial biases in the JPE ratings data from one state,[9] that this issue gained momentum. Other research confirms that in JPE surveys based

Footnotes

1. *See, e.g.*, American Bar Association, Guidelines for the Evaluation of Judicial Performance (1985); American Bar Association, Black Letter Guidelines for the Evaluation of Judicial Performance (2005), *available at* http://www.americanbar.org/content/dam/aba/migrated/jd/lawyersconf/pdf/jpec_final.authcheckdam.pdf [hereinafter ABA, Black Letter]; Seth S. Andersen, *Judicial Retention Evaluation Programs*, 34 Loy. L.A. L. Rev. 1375 (2001); Institute for the Advancement of the American Legal System, *Judicial Performance Evaluation in the States*, http://iaals.du.edu/initiatives/qualityjudges initiative/implementation/judicial-performance-evaluation (last visited Apr. 12, 2012).
2. *See* Institute for the Advancement of the American Legal System, *supra* note 1.
3. The first JPE program was started in Chicago in 1873, conducted by the Bar Association: "As the bar began to organize in order to combat the dominant role of partisan politics, surveys of lawyers were instituted to maximize the influence of the legal community on judicial selection." James H. Guterman and Errol E. Meidinger, In the Opinion of the Bar: A National Survey of Bar Polling Practices: A Research Project of the American Judicature Society (1977).
4. ABA, Black Letter, *supra* note 1.
5. *See* David C. Brody, *ABA Lawyers' Conference Model Survey Instruments*, *available at* http://www.americanbar.org/groups/judicial/conferences/lawyers_conference/resources/judicial_performance_resources.html [hereinafter Brody, *Model Survey Instruments*].
6. *See* Jennifer K. Elek, David B. Rottman & Brian L. Cutler, *Judicial Performance Evaluation Steps to Improve Survey Process and Measurement,* 96 Judicature 65 (2012).
7. *See* Christine Durham, *Gender and Professional Identity: Unexplored Issues in Judicial Performance Evaluation* Judges' J., Spring 2000, at 11 (2000); Justice D.K. Malcolm, Report of Chief Justice's Taskforce on Gender Bias (1994); Joyce S. Sterling, *The Impact of Gender Bias on Judging: Survey of Attitudes Toward Women Judges*, 22 Colo. Law Rev. 257 (1993).
8. *See* David C. Brody, *The Use of Judicial Performance Evaluation to Enhance Judicial Accountability, Judicial Independence, and Public Trust,* 86 Denv. U. L. Rev. 115 (2008).
9. Rebecca D. Gill, Sylvia R. Lazos & Mallory M. Waters, *Are Judicial Performance Evaluations Fair to Women and Minorities? A Cautionary Tale from Clark County, Nevada*, 45 Law & Soc'y Rev.731 (2011) [hereinafter Gill et al., *Women and Minorities*].

on the ABA model,[10] women and minority judges receive systematically poorer ratings on average relative to their male and majority group peers.[11]

Although critics have recently targeted the ABA model as a biased approach to performance evaluation,[12] stereotypic bias is likely present in the rating results of JPE surveys developed independently from this model. Gender and racial biases have been observed in both informal performance appraisal and formal work-performance evaluations across a number of job types and fields. When rating others' work performance, evaluators often draw on assumptions about race, ethnicity, gender, and other social or cultural stereotypes to construct their judgments. An evaluator may or may not be consciously aware of doing this. This cognitive phenomenon, known to some in the court community as *implicit bias*, has been found to produce systematically different judgments about candidates with identical qualifications or about employees with comparable performance to the systematic disadvantage of women and racial minorities subjected to evaluation.

Stereotypic bias is also likely to be present in the ratings from all JPE surveys because of how these surveys are developed. Many modern approaches to survey design improve the overall quality of survey data in part by reducing the likelihood or the impact of an array of undesirable response biases, including, in some cases, stereotypic gender and racial biases.[13] Most state JPE surveys do not reflect scientific advances in the understanding of quality survey design.[14] This is perhaps to be expected because most were designed a decade or more ago, generally by groups of legal practitioners with limited guidance, if any, from researchers with professional expertise in survey design.[15]

Efforts to incorporate modern techniques in survey design can help to minimize the impact of stereotypic biases on data and, more generally, can improve the overall quality of data collected by JPE surveys.[16] In this article, we draw on the experience of the first state to redesign its JPE surveys following the new research findings about systematic bias in JPE results.

IMPROVING SURVEY-BASED JUDICIAL-PERFORMANCE EVALUATION: A CASE EXAMPLE

In 2010, a Supreme Court of Illinois sought assistance from the National Center for State Courts (NCSC) in developing and implementing a new survey for use in their mandatory statewide JPE program. In this program, all judges are required to undergo evaluation for the purpose of judicial education and self-improvement. When selected, judges are asked to nominate attorneys and court personnel to complete their evaluations. They also meet with a facilitator or mentor judge at the end of the process to discuss individual results and to help create an action plan for professional development. Given the structure and goals of this program, the overseeing state Supreme Court committee prioritized the confidentiality of performance-evaluation results to encourage an open, honest atmosphere for feedback. They stressed the importance of confidentiality both for respondents providing feedback and for judges in terms of their individual results. Individual JPE results are not retained on file for any administrative purpose, nor are they shared with anyone but the judge and his or her facilitator.

Efforts to incorporate modern techniques in survey design can [minimize bias and] improve the overall quality of data . . . .

Within this framework, NCSC attempted to develop a new survey for this state JPE program that improved upon contemporary JPE survey practices. NCSC sought to do so, in particular, in ways designed to minimize the likelihood that the tool would produce systematically biased results as observed in other state JPE programs. In particular, we describe efforts undertaken to develop a new JPE survey instrument for use with attorney respondents. The new JPE survey instrument for statewide use emerged from the following multi-step process.

Critical review. NCSC staff conducted a review of 22 JPE surveys of attorneys to identify key judicial-performance criteria. This sample included four model surveys put forth by various organizations and a number of surveys recently or currently used by state JPE programs.[17] Informed by this review of existing survey instruments, NCSC staff then assembled a preliminary list of modified survey items that represented the criteria identified by the Illinois legal community as critical to judicial performance. Survey-design considerations at this stage emphasized basic item and response-scale clarity and correspondence, which many contemporary JPE surveys lacked. To reduce biased responding, NCSC focused particularly on developing items that described more concretely the kinds of judicial behaviors that an attorney or court staff would actually have the opportunity to directly observe. Similarly, questions that asked respondents to make generalized attribu-

10. *See* ABA, BLACK LETTER, *supra* note 1; *see also* Brody, *Model Survey Instruments*, *supra* note 5.
11. *See, e.g.*, Gary K. Burger, *Attorney's Ratings of Judges: 1998-2006*, MOUND CITY, MO: REPORT TO THE MOUND CITY BAR (2007); Rebecca Gill, *Judicial Performance Evaluations as Biased and Invalid Measures: Why the ABA Guidelines Are Not Good Enough* (2012), *available at* http://ssrn.com/abstract=2031800 [hereinafter Gill, *Biased and Invalid Measures*]; and NATALIE KNOWLTON & MALIA REDDICK, LEVELING THE PLAYING FIELD: GENDER, ETHNICITY, AND JUDICIAL PERFORMANCE EVALUATION (2012).
12. *See* Gill et al., *Women and Minorities*, *supra* note 9; Gill, *Biased and Invalid Measures*, *supra* note 11.
13. *See, e.g.*, Boris B. Baltes, C.B. Bauer & Peter Frensch, *Does a Structured Free Recall Intervention Reduce the Effect of Stereotypes on Performance Ratings and by What Cognitive Mechanism?* 92 J. OF APPLIED PSYCHOL. 151 (2007); Cara C. Bauer & Boris B. Baltes, *Reducing the Effects of Gender Stereotypes on Performance Evaluations,* 47 SEX ROLES 465 (2002); E. Lee Bernick & David J. Pratto, *A Behavior-Based Evaluation Instrument for Judges*, 18 JUST. SYS. J. 173 (1995).
14. *See* Elek et al., *supra* note 6.
15. *See, e.g.*, Steven Flanders, *Evaluating the Judges: How Should the Bar Do It?* 61 JUDICATURE 304 (1978); Gill et al., *supra* note 9; and GUTERMAN & MEIDINGER, *supra* note 3.
16. *See* Elek et al., *supra* note 6.
17. *See* Elek et al., *supra* note 6.

External testing is a critical step prior to full-scale implementation

• • • •

tions about the judge's performance or conjecture about the judge's personality were recast into more concrete, behavioral terms or eliminated from consideration.

Supreme Court committee oversight. The state Judicial Performance Evaluation committee consisting of judges and attorney members reviewed preliminary drafts of the survey instrument and provided recommendations for revision of content. These reviews helped to ensure that survey items captured the key elements of judicial performance as defined in Illinois. Committee feedback included comments on the evaluation criteria represented in the survey items and the language or legal terminology used. This committee also reviewed and commented on several subsequent drafts of the instrument as it was refined in the following steps.

We wish to note here that other states seeking to develop a new JPE program may also benefit from conducting separate focus groups of judges and of each potential respondent group (*e.g.*, attorneys, court personnel). Focus groups, when conducted by independent research groups using trained, professional facilitators, can yield rich information and honest, constructive input about the types of information judges are most interested in learning and that they would find most helpful, and about the types of observations that respondents feel they ought to be able to communicate in a constructive evaluation of judicial performance. Through this approach, stakeholder concerns may be addressed at an early stage of program development. These outreach efforts may also help to promote the upcoming program and generate support. By engaging stakeholders in the development process, judges, attorneys, and others involved can develop a sense of ownership over the program, leading to greater satisfaction with the final product. This may be important for some types of JPE programs more than others (*e.g.*, for those designed for the purpose of professional development or voter education).

Consultation with survey design and work-performance evaluation experts. NCSC staff also consulted with academic experts on performance evaluation and survey design to further improve evaluation accuracy and minimize the opportunity for systematic biases based on gender, race, or ethnicity to influence evaluation responses. The draft survey was refined based on feedback from this panel of experts to include more concrete language in the description of survey items. A *structured free-recall task*, in which survey respondents are prompted to recall specific instances of the judge's actual courtroom behavior immediately before completing the judge's performance evaluation, was also adopted based on expert recommendations. Research shows that this type of task facilitates retrieval of information about past observed behavior for use in the formulation of performance-evaluation judgments, reducing reliance on social schemas (*e.g.*, stereotypes). This helps to produce less systematically biased and more accurate evaluations.[18]

Testing. Following these steps and in preparation for full-scale launch of the JPE survey, NCSC staff created the survey in a web-based environment using the Confirmit software platform with methodology that comported with Dillman's scientific tailored design method for internet surveys.[19] This approach includes a research-informed procedure for scheduling and issuing tailored notifications according to the respondent's status (*i.e.*, if the survey is complete, incomplete, or not yet started). Notifications designed for the Illinois JPE survey include a prenotice in which the respondent is notified of his or her selection for participation before the evaluation period, an invitation at the beginning of the evaluation period, and up to three tailored reminder notices, which may be issued to the respondent until he or she participates in the evaluation or until the evaluation period concludes. [20]

After the JPE survey tool was developed in the web-based environment, NCSC staff conducted a careful internal test to ensure that the mechanics of the internet survey and corresponding distribution processes operated as intended. After passing this internal test of general functionality, the survey was subjected to external testing with samples of eligible respondents. External testing is a critical step before full-scale implementation that can help establish instrument validity and determine whether efforts to minimize or eliminate systematic biases in results were successful. In Illinois, two external tests were conducted.

First, NCSC staff contracted with a local research agency to evaluate the JPE survey by conducting *cognitive interviews* with three licensed Illinois attorneys. In this cognitive-interview approach, attorneys completed the online evaluation form in the presence of interviewers who were trained to assess problems with survey items, instructions, and functionality in this context based on Tourangeau's cognitive-interviewing model.[21] Interviewers asked attorneys probing questions about their thought processes and reactions as they completed the survey to determine which components, if any, presented barriers to participation. This included probes to identify components that lacked clarity, did not use appropriate legal terminology, were unnecessarily long or tedious, or posed other challenges to respondents. These trained interviewers identified user concerns regarding the clarity of some instructions (*e.g.*, the explanation of the con-

18. *See e.g.*, Bauer & Baltes, *supra* note 3, and Baltes et al., *supra* note 3.
19. *See* Don A. Dillman, Jolene D. Smyth & Leah Melani Christian, Internet, Mail, And Mixed-Mode Surveys: The Tailored Design Method (3d ed. 2008).
20. The Illinois surveys routinely achieve response rates in the range of 55-70%, with approximately 55-60% of attorneys and approximately 60-70% of court personnel completing the JPE surveys in full. The use of reminder notifications was associated with an increase in response rates of 25 percentage points. Note also that respondents who preferred to complete a hard copy of the survey were provided with this alternative to participate.
21. Roger Tourangeau, *Cognitive Science and Survey Methods: A Cognitive Perspective, in* Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines (Thomas B. Jabine, Miron L. Straf, Judith M. Tanur & Roger Tourangeau eds., 1984); and Gordon B. Willis, *Cognitive Interviewing: A "How To" Guide* (1999), *available at* http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf.

fidentiality policy), the user-friendliness of some components of survey navigation, and the clarity of some survey items.

In addition to cognitive-interview testing, NCSC staff conducted a pilot study of the JPE survey to vet the JPE survey instrument and procedure. A small sample of judges volunteered to participate in this pilot study, which produced complete survey data from a sample of approximately 100 eligible attorney respondents. These pilot study respondents were also asked to complete an optional follow-up questionnaire designed to elicit feedback about respondent perceptions of and experience with the online JPE survey tool. Based on statistical analysis of this JPE survey pilot data, user feedback from the follow-up questionnaire, and results from the cognitive interviews, instructions were refined and streamlined, and problematic items were revised or removed to improve overall clarity, user-friendliness, reliability, and validity of the JPE survey.

The new evaluation instrument that emerged from this multi-step development process contained 59 rating questions and five optional narrative comment fields across the following five areas of judicial performance: legal and reasoning ability, impartiality, professionalism, communication skills, and management skills.[22] The instrument met psychometric standards and adhered to best practices in survey design and performance evaluation, with a particular focus on minimizing the potential for an array of respondent biases (including stereotypic biases). NCSC staff also adopted procedures to enhance data quality control within the framework of the existing state JPE program. First, respondents were assigned individual logins to access the JPE survey; respondents could therefore complete an evaluation of a single judge only once within a single evaluation period. Respondents were also prompted to base their evaluations on their own recent, direct experience working with the judge in a workplace environment, and not on the judge's reputation or on personal or social contact with the judge. By incorporating the structured free-recall task discussed above[23] into the web-based JPE survey, respondents were explicitly prompted to recall their direct experiences working with the judge before completing the judge's evaluation. With these efforts, authors hoped to facilitate respondent use of more reliable sources of information about each judge's performance in the evaluation process.

The present study illustrates one potential approach to the development of a fairer JPE survey tool. An analysis of the first full year of data produced by the Illinois survey revealed that JPE results did not systematically differ by the judge's gender.[24] This demonstrates that a JPE survey can be developed that both comports with the conceptual underpinnings of the influential ABA model and produces results without marked gender disparities, as has been found in the results of JPE surveys done elsewhere.[25]

While we recommend a rigorous development process like the one used in Illinois, the survey that emerged from that development process should not be unquestioningly adopted by other states. Several important features of the state context for JPE should be carefully considered when approaching survey redesign or the development of a new tool, including the expressed purpose of the JPE program (*e.g.*, to inform the individual judge's professional development, to inform the assignment and/or retention decisions of the judiciary, to inform the public) and the state's distinctive legal and judicial culture. A need may arise to develop separate JPE surveys to evaluate judges presiding over different case types or dockets. For example, a few states have already developed separate survey-evaluation processes for use with judges presiding in high-volume and low-volume courts. The real challenge for the future, however, is to develop multi-method JPE programs that call upon a diverse set of data-collection strategies to maximize evaluation accuracy and utility.

[R]ecent empirical findings support a policy recommendation that calls for the validation and likely revision of [JPE] survey instruments . . .

BEYOND SURVEY-BASED MEASURES

Taken together, recent empirical findings support a policy recommendation that calls for the validation and likely revision of survey instruments employed by JPE programs, and for additional guidance on multi-method approaches to the measurement of judicial performance. To date, the over-reliance on surveys has been a significant a problem because the weaknesses of surveys are not compensated for by the strengths of alternative measurement methodologies. A future program of research should explore how other methodologies may help to enhance the quality of JPE programs and improve the ability to achieve expressed JPE goals.

The quest for a better, multi-method program of JPE is a complicated one. Commonly recognized objective measures of judicial performance tend to emphasize productivity over quality. More subjective forms of evaluation like survey ratings, narrative feedback, and courtroom observation tend to be relied upon to capture performance quality. It should be recognized that all subjective forms of evaluation are capable of producing biased results if they are not designed well, as has been observed with survey-based measures. Greater structure is likely needed to establish a sound process for evaluators. These forms of JPE should also be subjected to scientific scrutiny before they are adopted to ensure that they produce fair, high-quality evaluation data.

Efforts to revitalize and improve JPE programs are already

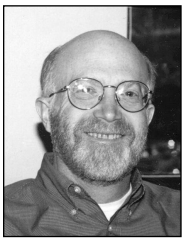
22. The Illinois survey has not been published in final form, but a draft version was included as an appendix to Knowlton & Reddick, *supra* note 11.
23. Bauer & Baltes, *supra* note 3; and Baltes et al., *supra* note 3.
24. The Administrative Office of the Illinois Courts does not track gender or racial demographic information for judges statewide. Researchers could determine gender based on the name of the evaluated judge, but racial background could not be readily identified. Thus researchers could not evaluate Year 1 results from the new JPE survey to determine whether systematic racial bias was present. However, authors expect results from an analysis for racial bias to be similar to the results of the gender analysis because of the similar psychological mechanisms underlying biased responding.
25. *Cf.*, Burger, *supra* note 11; Gill, *supra* note 11; Gill et al., *supra* note 9.

underway. Several other states have followed Illinois's lead and are overhauling JPE programs that have been in place for 10 years or more. The international scene also holds potential as we look for other approaches to evaluating judges. Most Western European countries have JPE programs in place, some with decades of experience.[26] To promote an international dialogue on JPE, the National Center for State Courts and the Academy of the Social Sciences in Australia co-organized a Workshop on Evaluating Judicial Performance, bringing together an international group of 22 judges, law professors, and social scientists. The workshop, held May 9-10, 2013, at the International Institute for the Sociology of Law in Oñati, Spain, identified issues that can be regarded as generic to the task of evaluating judges, along with the key differences associated with distinctive court structures, legal systems, and, most importantly, recruitment to the bench.[27]

CONCLUSION

Judicial-performance evaluation programs can be of great benefit to the state courts. They can help to address core concerns, such as the need to be accountable in ways consistent with judicial independence and to allow judges the opportunity to hone their skills on the bench. The potential contributions of JPE programs remain, but considerable work is needed to bring JPE surveys up to the standard of best practices and to balance the results of those surveys against other well-developed approaches to performance evaluation. Although some states already have programs that are multi-method (augmenting surveys with interviews, case-processing data, and other measures), the problem of bias will need to be tackled. With states working to improve their JPE programs and international attention growing in this area, it is likely that JPE programs of the future will look very different than they do today.

Jennifer Elek, Ph.D., is a court research associate with the National Center for State Courts. Dr. Elek's work at NCSC includes projects on judicial performance evaluation, problem-solving courts, offender risk and needs assessment, and gender, racial, and ethnic fairness in the courts. Dr. Elek holds a Ph.D. in social psychology with a concentration in quantitative methods from Ohio University, an M.A. from the College of William and Mary, and a B.A. from Vassar College.



David B. Rottman, Ph.D., is a principal court research consultant at the National Center for State Courts. His current responsibilities include directing projects that evaluate Brooklyn's Red Hook Community Justice Center, improve judicial performance evaluation programs, and develop new models for organizing state courts based on the Kennedy School of Government/NCSC "Executive Session on State Court Leadership in the 21st Century." He is a cofounder with Judge Kevin Burke, Judge Steve Leben, and Professor Tom Tyler of www.proceduralfairness.org, which promotes the application of procedural-fairness principles in court reform.

26. Recruitment, Professional Evaluation and Career of Judges and Prosecutors in Europe: Austria, France, Germany, Italy, the Netherlands and Spain (Giuseppe Di Federico ed., 2005).
27. For information about the workshop, visit http://goo.gl/tmlaT.