Oñati International Institute for the Sociology of Law
Instituto Internacional de Sociología Jurídica de Oñati
Oñati Socio-Legal Series
ISSN: 2079-5971

Oñati Socio-legal Series, v. 4, n. 5 (2014) – Evaluating Judicial Performance
ISSN: 2079-5971

Methodologies for Measuring Judicial Performance: The Problem of Bias

JENNIFER K. ELEK*
DAVID B. ROTTMAN*

Elek, J.K., Rottman, D.B., 2014. Methodologies for Measuring Judicial Performance: The Problem of Bias. *Oñati Socio-legal Series* [online], 4 (5), 863-879. Available from: http://ssrn.com/abstract=2533937

Abstract

Concerns about gender and racial bias in the survey-based evaluations of judicial performance common in the United States have persisted for decades. Consistent with a large body of basic research in the psychological sciences, recent studies confirm that the results from these JPE surveys are systematically biased against women and minority judges. In this paper, we explain the insidious manner in which performance evaluations may be biased, describe some techniques that may help to reduce expressions of bias in judicial performance evaluation surveys, and discuss the potential problem such biases may pose in other common methods of performance evaluation used in the United States and elsewhere. We conclude by highlighting the potential adverse consequences of judicial performance evaluation programs that rely on biased measurements.

Key words

Judicial Performance Evaluation; judges; measurement; prejudice and discrimination

---

Article resulting from the paper presented at the workshop *Evaluating Judicial Performance* held in the International Institute for the Sociology of Law, Oñati, Spain, 9-10 May 2013, and coordinated by Francesco Contini (National Research Council of Italy), Jennifer Elek (National Center for State Courts), Kathy Mack (Flinders University), Sharyn Roach Anleu (Flinders University) and David Rottman (National Center for State Courts).
We are grateful to Jordan Bowman for his research and editorial skills.

* Jennifer K. Elek is a court research associate at the National Center for State Courts. In Dr. Elek's recent work at NCSC, she has focused on promoting gender and racial fairness in the courts; on improving judicial performance evaluation programs in the United States; on educating the court community about offender risk and needs assessment and its role in evidence-based sentencing; and on identifying and evaluating the efficacy of problem-solving court programs. She holds a Ph.D. in social psychology from Ohio University, an M.A. from the College of William and Mary, and a B.A. from Vassar College. National Center for State Courts. 300 Newport Avenue, Williamsburg, VA 23185. jelek@ncsc.org
* David B. Rottman is a principal court researcher at the National Center for State Courts. His current research concerns the effectiveness of specialized courts, minority group opinions of courts, and the efficacy of surveys for evaluating judicial performance. With Tom Tyler and Judges Kevin Burke and Steve Leven, he co-founded www.proceduralfairness.org to promote implementation of procedural justice principles to all aspects of court operations. He previously worked at the Economic and Social Research Institute in Dublin, Ireland. He holds a Ph.D. in sociology from the University of Illinois at Urbana. National Center for State Courts. 300 Newport Avenue, Williamsburg, VA 23185. drottman@ncsc.org

Resumen

Durante décadas ha habido una preocupación por la discriminación por género y racial en las evaluaciones del rendimiento judicial basadas en encuestas, comunes en Estados Unidos. De acuerdo con un gran corpus de investigación básica en las ciencias psicológicas, estudios recientes confirman que los resultados de estas encuestas de evaluación del rendimiento judicial están sistemáticamente sesgados contra las mujeres y los jueces de minorías. En este artículo se explica la manera insidiosa en que las evaluaciones de rendimiento pueden estar sesgadas, se describen algunas técnicas que pueden ayudar a reducir las expresiones de sesgo en los estudios de evaluación del rendimiento judicial, y se debate el problema potencial que estos sesgos pueden plantear en otros métodos comunes de evaluación del rendimiento utilizados en Estados Unidos y otros países. Se concluye destacando las posibles consecuencias adversas de los programas de evaluación del rendimiento judicial que se basan en mediciones sesgadas.

Palabras clave

Evaluación del rendimiento judicial; jueces; medición; prejuicio y discriminación

Table of contents

1. Introduction: the problem

Judiciaries in many common law and bureaucratic legal systems have developed methodologies for evaluating the performance of their judges. There is insufficient recognition that the choice of evaluation methods can have unintended consequences for the composition of the judiciary, favoring some skills and qualities over others, and favoring some groups of judges over others. The problem this presents is not unique to the evaluation of judicial performance, nor is it limited to only certain types of evaluation methodologies. Whether a performance evaluation program draws on information from surveys, interviews, quality reviews, or observations, the very process of evaluating job performance of any kind may be systematically biased against certain demographic groups. When rating others' work performance, evaluators are likely to draw on assumptions about race, ethnicity, gender, and other social or cultural stereotypes that ultimately produce systematically biased assessments of performances (e.g., Deaux and Emswiller 1974, Martell 1996, Biernat *et al*. 2012).

This article focuses on the problem of bias in judicial performance evaluation programs in the United States, although many of the observations and conclusions offered are likely to be of more general application. There are several reasons for our limited focus. One is that the United States presents a variety of approaches to judicial evaluation rather than a clear national model. A second reason is that the available research regarding systematic gender and racial bias in evaluations of judges is almost entirely drawn from the United States. The precise nature of these biases may differ between cultures, but the mechanisms through which they operate to influence judgment are likely universal.

We have three main objectives for this article. First, we explain the insidious manner in which assessments of job performance, generally, and judicial performance, specifically, may be biased. Second, we discuss the possibility of reducing expressions of bias in the most common methodological approach to measuring judicial performance in the United States. Third, we begin to explore the opportunities for bias in other common methods of performance evaluation and examine some practical concerns associated with the problem of bias in judicial performance measures. We begin, however, by describing the context in which judicial performance evaluation programs operate and consequences of these programs in the United States.

2. Evaluating judges in the United States

Official judicial performance evaluation (JPE) programs operate in 18 states plus the District of Columbia (Strickland *et al*. 2012, Table 6; Knowlton and Reddick 2012). Until recently, most JPE programs served only the trial bench, but a few states now also conduct evaluations of appellate judges. Evaluation results are typically prepared for and distributed among any of three main audiences: individual judges, court administration, and the public.

In practice, states often use the same JPE results to serve several purposes simultaneously. First, JPE results may be provided to individual judges on a confidential basis to assist in self-improvement (six states).[1] The results may not be included in any official record, but may be shared and discussed with a facilitator or mentor judge. Second, JPE results are used internally to inform administrative decision-making in states in which judges are appointed by the governor (three states). JPE results in those states become a part of a judge's personnel record, to be used at fixed intervals to decide whether a judge should be reappointed and the nature of his or her court assignment.[2] These states are clustered in the New

[1] Florida, Hawaii, Idaho, Illinois, Massachusetts, New Hampshire, and Rhode Island (in Hawaii and New Hampshire aggregate survey results are made public to enhance public trust in the courts).
[2] Connecticut, New Jersey, and Vermont (and the District of Columbia).

Oñati Socio-legal Series, v. 4, n. 5 (2014), 863-879
ISSN: 2079-5971

England region of the country and are characterized by strong administrative judges with significantly more control over judicial careers than is typical in the rest of the United States. Third, JPE results may be provided to the public for informational purposes to educate citizens and enhance public confidence in the courts (two states). This information may be published in summary format, without identifying individual judicial performance evaluation results. Fourth, JPE results are a part of the information provided to judicial performance evaluation commissions and the voting public in states following the Missouri Plan of judicial selection to inform judicial retention decisions (eight states).[3] In these states, judicial nomination commissions recommend applicants for judicial office to the governor, who must appoint a candidate from the supplied list. Subsequently, at fixed intervals, sitting judges run against their own record for retention in office. Performance evaluation commissions in those states use JPE results to inform recommendations to the voting public regarding whether or not to retain these judges. To inform voters prior to the retention election, the commissions publish their recommendations along with the detailed JPE results and supplementary information about each judge. Often, states which use JPE results to inform personnel decisions also distribute results to judges to inform self-improvement and to the public in some capacity to promote transparency objectives.

All types of JPE programs in the United States rely significantly on survey methods of data collection, surveying practicing attorneys and others for feedback about a judge's performance on the bench. This practice dates back to 1873, when a Chicago area bar association distributed surveys to attorney members and packaged the results for distribution to voters on or immediately before election day. This first (unofficial) JPE program in the United States served a political purpose. It was a reaction by the organized bar to the growing practice of partisan judicial elections, which diminished the role of the legal profession in determining the composition of the bench: "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" (Guterman and Meidinger 1977). The use of such surveys persists today in many large cities as unofficial programs.[4] When the first official state JPE programs emerged in the mid-1970s, the reliance on surveys of lawyers was maintained, augmented in some instances to include surveys of court clerks and other courtroom personnel, and of jurors and litigants as they left the courtroom.

The diverse practices and designs of JPE programs became a concern of the American Bar Association (ABA), which assumed a national leadership role in formalizing and improving the evaluation process. In 1985, the ABA first established model standards for conducting JPE programs, called the ABA Black Letter Guidelines on Judicial Performance Evaluation. The ABA model identified the appropriate goals and uses of these programs, described proper administration and dissemination practices, and identified the criteria on which they recommended that judges be evaluated. In 2005, the ABA Judicial Division's Lawyers Conference, in collaboration with the ABA Justice Center Standing Committee on Judicial

[3] Alaska, Arizona, Colorado, Kansas, Missouri, New Mexico, Tennessee, and Utah.

[4] Official or unofficial, JPE results have great power to dramatically influence the careers of judges. These evaluation results can factor into judicial nominations and into judicial election campaigns as ammunition by opposing sides, even when judicial elections are normally non-partisan. For example, members of the University of Chicago Law and Economics faculty became interested in measuring judicial quality by proposing a "tournament of judges" methodology in which all federal appellate judges would be ranked to determine the best judge in the United States. While acknowledging the limitations of their measures of quality, the raters have not hesitated to interject these rankings into policy issues, most notoriously using their rankings to raise doubts about Justice Sonia Sotomayor's fitness for the U.S. Supreme Court during her confirmation hearings (Levy *et al*. 2010, pp. 321-322).

Independence, revised the model guidelines[5] and, as a follow up, developed survey instruments that they promoted as models for general use.[6]

Although the ABA model achieved wide acceptance, the performance evaluation surveys it inspired became a source of discontent among some judges and lawyers. Early surveys were developed largely by committees of judges and lawyers who applied adaptations of the verbatim ABA black letter guidelines to survey form, with limited participation by political scientists or survey design experts. Even the model surveys promoted by the ABA Lawyers Conference as a supplement to the 2005 revision of the ABA model failed to adequately take into account current best practices in survey design and, in particular, developments related to measurement of work performance (see Elek *et al*. 2012). Judges participating in one state JPE program expressed broad skepticism about the quality and validity of data drawn from their state survey, indicating that judges themselves may recognize the problems associated with these traditional JPE instruments (Institute for the Advancement of the American Legal System 2008).[7]

In addition to fundamental problems of design, some court professionals raised early concerns that JPE surveys based on the ABA model produced results that were systematically biased against women and minority judges (Malcolm 1994, Durham 2000, Burger 2007, see also Resnik 1996, Kearney and Sellers 1996). These early voices included the social scientist responsible for designing the surveys still being used by the Colorado judiciary, who found "clear bias related to gender in lawyer evaluations of judges, with female judges ranked lower in all attributes measures to a statistically significant degree" (Sterling 1993). This bias apparently transferred to the JPE Commission's recommendation on retention, in which 15 percent of male judges but 25 percent of female judges received a recommendation for "no retention". Other evidence of gender bias among attorneys in their assessments of judges started to emerge as early as the 1980s through the findings of gender bias task forces convened at the state or national level. Justice Christine Durham (2000), for example, cited the conclusion of an ABA Commission on Women in the Profession: "Even women who enjoy the prestige of the judiciary are affected by bias. Judicial evaluation programs reflect that women judges endure consistently stronger criticism than their male colleagues, especially in subjective categories such as demeanor." Despite these early warnings about gender and racial bias in survey-based JPE results, state JPE programs continued to rely heavily the survey method, with few if any revisions to the instrumentation used.

Only more recently have the state courts started to take notice of the problem of systematic racial and gender bias in survey-based JPE programs. An article published in *Law & Society Review* demonstrated the presence of such biases in the JPE data from one state (Gill *et al*. 2011). The primary author of that article has become an outspoken opponent of the popular ABA model and has cited a number of studies of survey-based ABA-style JPE programs (e.g., see Burger 2007, Knowlton and Reddick 2012) and her own analyses of state JPE data (see Gill *et al*. 2011, Gill 2012, Gill and Retzl 2013) to criticize the model as biased against female and minority judges. As many state-run JPE programs are predicated on the ABA model (Gill 2012) and use these JPE results to inform an array of judicial assignment, retention (via both popular election and administrative review), education, and self-improvement decisions, such findings raise fundamental

[5] The guidelines are available at http://www.americanbar.org/content/dam/aba/publications/judicial_division/jpec_final.authcheckdam.pdf.

[6] The model surveys are available online at http://www.americanbar.org/groups/judicial/conferences/lawyers_conference/resources/judicial_performance_resources.html.

[7] In this 2008 survey, Colorado judges were asked if the "validity and accuracy of survey responses" was "not a major problem," a "minor problem," or a "major problem." Only 13 percent indicated that survey validity was not a major problem.

questions about the validity of existing JPE surveys. As court leadership looks to revitalize their state JPE programs, the question is now not one of whether gender and racial biases may be present in JPE data, but how the general problem may be effectively addressed.

3. Bias in perception and judgment: lessons from psychological research

Lessons learned from subfields of psychology offer critical insights to the discussion of appropriate methodologies for evaluating the performance of judges who sit on the bench. Basic research in the fields of social psychology and social cognition examines the cognitive processes underlying social perception and judgment, including the limitations of human judgment and the conditions that may help to minimize expressions of bias. Researchers in the field of industrial/organizational (I/O) psychology work to improve organizational effectiveness by better understanding how employee-level characteristics interact with features of the workplace environment, organizational culture, and management system. A specialized area of research in I/O psychology focuses on effective job performance evaluation (also referred to as job appraisal) techniques and performance management processes. The research emerging from these two fields sheds additional light on questions about how stereotypic biases may influence performance evaluations and how the nature of the evaluation may serve to exacerbate or minimize the effects of such biases on judgment. These general findings have significant implications for many judicial performance evaluation programs currently operating in the United States.

This body of research has confirmed that normative views of most professional occupations include assumptions about specific skills or personal traits necessary to excel on the job, and that these job-specific traits or characteristics can also be linked to learned stereotypes about a particular social group (e.g., Biernat and Kobrynowicz 1997). In general among Americans, communal traits (e.g., warm, helpful, kind, nurturant, gentle, interpersonally sensitive) tend to be automatically associated with women, whereas agentic traits (e.g., competent, dominant, independent, ambitious, confident, prone to act as a leader) tend to be automatically associated with men (Rudman and Glick 2001, see also Eagly and Karau 2002). Similarly, occupations perceived as primarily lower-status and communal in nature (e.g., secretary, nurse) tend to be associated with women, and occupations perceived as primarily higher-status and agentic in nature (e.g., bank vice president, physician) tend to be associated with men (Eagly *et al*. 2000). Indeed, a marked gender bias exists in the legal profession: even the latest generation of law students automatically associates the profession of judging with men and not with women, linking women to the home and family (Levinson and Young 2010). Such associations do not necessarily reflect attitudes that are consciously endorsed, but instead reflect learned information (such as cultural stereotypes) about gender and race that can, without awareness or conscious effort, inform social perception, judgment, and/or behavior towards others. These *implicit* associations may contribute to the development of stereotypic expectations for on-the-job work performance that can color assessments of a particular candidate's actual qualifications or a particular employee's actual performance behavior.

When performance-related judgments require respondents to ascribe personality characteristics to an individual or develop higher-order attributions about the individual's traits or abilities, these judgments may in turn be informed by stereotypes (see Dunning and Sherman 1997). One study illustrates this phenomenon well: American participants tended to explain a male's successful performance on a masculine sex-typed task as indicative of his ability, but tended to attribute a female's successful performance on the same task to luck (Deaux and Emswiller 1974). It appears to be more difficult for Americans to infer agentic traits than communal traits from behavior when the performer is female than when the

performer is male (Scott and Brown 2006). In addition, stereotypes may subtly alter the standards used in evaluation if the provided standards are poorly defined. In one empirical study in a hiring context, people asked to evaluate a female and a male candidate for a masculine sex-typed police chief position rated whichever qualification (street smarts, formal education) the male applicant possessed but the female applicant did not as the more important qualification for the job (Uhlmann and Cohen 2005). A similar "shifting standards" problem may occur in on-the-job performance evaluation (c.f. Biernat *et al*. 1998). Because these subtle stereotypic associations may operate "behind the scenes" to make cognitive processing easier and more efficient for the individual, even those who believe themselves to be egalitarian may find themselves inadvertently attributing different causes for male versus female employee performance, or using slightly different performance measurement standards in their evaluations of employees from different social groups when the evaluation is not constructed in ways that help to minimize these problems.

Stereotypic biases may emerge not only in performance evaluations provided by members of the socially dominant group, but also from members of the stigmatized group(s). For example, the creator of a popular test of implicit racial bias (i.e., a form of racial bias that can operate below the level of conscious awareness and that can operate even if prejudicial attitudes are not personally endorsed) has published data showing that a substantial proportion of African Americans who have completed the test implicitly favor the white majority (Greenwald and Krieger 2006). Thus even members of groups that are themselves disadvantaged by cultural stereotypes can acquire implicit associations based on them. Because anyone may develop these biased implicit associations, the implication is that regardless of the evaluator's identity or demographic background, the potential to evaluate the performance of others in a stereotypically biased manner exists.

Even relevant training and expertise do not universally inoculate an individual against the potential for stereotypical bias to influence judgment. Judges are susceptible to an array of common cognitive biases in decision-making (e.g., Guthrie *et al*. 2001, Englich *et al*. 2006), although they may be somewhat better than other experts and laypersons at overcoming biases in certain situations relevant to their area of expertise if they are motivated to do so (e.g., Wistrich *et al*. 2005, Rachlinski *et al*. 2009). Perhaps in part because of their knowledge and expertise, legal professionals tend to overestimate their own ability to disregard extra-legal factors and render unbiased judgments. For example, 97 percent of judges in one study rated themselves in the top half of the group on their ability to "avoid racial prejudice in decisionmaking" (Rachlinski *et al*. 2009, p. 126). This is particularly noteworthy given evidence in the same study that white judges also exhibit implicit racial bias and that such bias appeared to influence their mock sentencing decisions. There is no easy solution for the general problem these subtle biases pose in professional judgment, but targeted strategies may help to minimize expressions of subtle biases when the proper resources are available to do so (for more, see Casey *et al*. 2012).

4. Recent efforts to address bias in survey-based JPE programs in the United States

Given the emerging empirical evidence of systematic bias in JPE data, some state courts and professional court organizations have begun to take remedial steps. For example, Administrative and judicial leaders in one American state (Illinois) contracted with the National Center for State Courts (NCSC) to redesign their JPE survey instrument and associated methodology when the state Supreme Court ruled to change the survey-based JPE program from voluntary (see Cermak and Block 2001) to mandatory for all of the state's trial court judges for the purpose of judicial education and professional development. Reducing the potential for systematic bias in the survey-based ratings was an explicit goal of the project.

Oñati Socio-legal Series, v. 4, n. 5 (2014), 863-879
ISSN: 2079-5971

NCSC staff worked with court leadership in Illinois to develop a new survey for this state JPE program that improved upon contemporary JPE survey practices while working within the program structure authorized by the Supreme Court. In particular, NCSC sought to design the survey in such a way as to minimize the likelihood that the tool would produce the kind of systematically biased results against female and minority group judges observed in other JPE programs.

The new Illinois JPE survey instrument emerged from a multi-step process (see Elek and Rottman 2013). The process included a thorough review of current approaches to JPE to inform preliminary survey development efforts.[8] Working closely with a dedicated Illinois Supreme Court Judicial Performance Evaluation Committee comprised of judges and attorney members as well as representatives from the Administrative Office of the Courts, NCSC staff then developed and refined a list of survey items that represented the criteria identified by the Illinois legal community as critical to judicial performance. These criteria mirrored those promoted by the ABA. Survey design considerations at this stage emphasized basic item and response scale clarity and correspondence, which many contemporary JPE surveys based on the ABA model lacked (Elek *et al*., 2012). To reduce biased responding, NCSC focused particularly on developing items that more concretely described the kinds of judicial behaviors that an attorney or that court staff would actually have the opportunity to directly observe. Similarly, questions that asked respondents to make generalized attributions about the judge's performance or conjecture about the judge's personality were recast into more concrete behavioral terms or eliminated from consideration. In addition, NCSC staff consulted with academic experts on performance evaluation and survey design to further improve the construction of the instrument in ways that would increase utility and accuracy while minimizing the opportunity for known response biases to systematically skew evaluation results. Expert input resulted in further refinements and the addition of one new major feature: A *structured free-recall task*, in which survey respondents are prompted to recall specific instances of the judge's actual courtroom behavior immediately prior to completing the judge's performance evaluation, was incorporated into the JPE survey procedure. This structured free-recall 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 (e.g., Bauer and Baltes 2002, Baltes *et al*. 2007). Based on this general review, close work with the JPE subcommittee and AOIC staff, and consultation with performance evaluation experts, a draft evaluation survey tool was developed.

In preparation for full-scale launch, NCSC staff created the new JPE survey in a web-based environment with methodology that comported with Dillman's scientific tailored design method for internet surveys (Dillman *et al*. 2009). This approach includes a research-informed procedure for scheduling and issuing tailored notifications according to the respondent's status (e.g., if the survey is complete, incomplete, or not yet started) to generate higher response rates. NCSC staff also adopted procedures to enhance data quality control within the framework of the existing state JPE program. This included login security measures to ensure that (a) only respondents with professional working experience with the judge could complete the evaluation survey and (b) respondents could submit 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

[8] NCSC staff reviewed twenty-two current or recently used state JPE survey tools and four model survey tools in this process. The development of the attorney version of the survey tool is described herein. Additionally, a version of the tool for use with court personnel respondents was also developed using similar techniques which produced similar outcomes.

task discussed above into the web-based JPE survey, respondents were explicitly prompted to recall their direct experiences working with the judge prior to completing the judge's evaluation. With these efforts and the efforts described above in the survey construction process, NCSC hoped to collect with the new web-based JPE survey more reliable data about each judge's actual performance.

The new web-based JPE survey tool was tested two ways with samples of eligible Illinois respondents. 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 (1984) cognitive interviewing model. In addition to cognitive interview testing, NCSC staff conducted a pilot study of the JPE survey to vet the JPE survey instrument and procedure using a small sample of five volunteer Illinois judges and 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 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 for full-scale implementation.

The new evaluation instrument emerged from this multi-step development process containing 59 rating questions and five optional narrative comment fields across the following five general content areas: legal and reasoning ability, impartiality, professionalism, communication skills, and management skills. Based on data from the first year of program operation, the instrument met psychometric standards for measurement reliability (Nunnaly 1978) for each of the five performance area subscales and for a total score index (computed as the average across each of the five subscales), $\alpha \geq .750$. All average inter-item correlations fell within the recommended range of .15-.50 (Clark and Watson 1995). Program records to date indicate that 55-65% of invited attorneys routinely complete the JPE survey in full, either via the web-based format or through an alternative hard copy submission option. The use of reminder notifications was associated with a response rate increase of 25 percentage points. This is significantly higher than attorney response rates reported in other JPE programs (e.g., 20%; see Brody 2008). Importantly, the survey instrument also produced JPE total and subscale results for male and female judges that did not significantly differ by gender, ts < 0.950. Future research to monitor and extend these findings, and to ensure similar equity between racial majority and minority judges, is strongly recommended.

The Illinois experience demonstrates that a more scientifically rigorous survey development process and the application of novel bias-reduction techniques can help to minimize the kind of systematically biased responding that has been observed in data from traditional JPE survey instruments in the United States. This is promising for other states, as several are also currently engaged in a reevaluation of their JPE programs and associated survey tools. Other professional organizations have called for states to minimize the potential impact of bias in survey-based judicial performance evaluation programs, with one group recommending the Illinois JPE survey process as a new model standard (Knowlton and Reddick 2012). In our view, however, much work remains to improve judicial performance evaluation measurement standards in the United States. Additional performance evaluation techniques developed by industrial/organizational psychologists and survey design experts should be considered for their potential value and feasibility based on the available resources and goals of each state JPE program. Other important factors, such as the state culture and general court

Oñati Socio-legal Series, v. 4, n. 5 (2014), 863-879
ISSN: 2079-5971

community attitudes toward JPE, may also influence JPE program development and design.

The potential for systematic bias is a necessary methodological consideration in the development or redevelopment of any contemporary JPE survey given the context in which these programs operate in the United States, even as the conceptualization of "good judging" evolves. Although the ABA model is the prevailing account of how judging is currently defined and evaluated in the United States, some states have already begun to incorporate other criteria into their survey-based JPE programs. These include concepts birthed from the field of psychology, such as procedural justice theory (Tyler 1990, 2007). Survey measures of perceptions of procedural fairness, however, have been shown to differ systematically depending on the gender of the person being evaluated (Johnson *et al*. 2007). This further supports the argument that any survey measure used in the evaluation of judges ought to be subjected to careful empirical scrutiny prior to full scale implementation.

5. The potential for bias in other common JPE measurement approaches

The survey method is useful when seeking to gather information from a large group of individuals in a short period of time at a relatively low expense and, if executed well, is likely to remain a staple of JPE programs in the United States. Our analysis leads us to recommend that JPE programs in the United States and elsewhere not only seek to refine their methods of survey-based JPE data collection, but also to complement survey-based JPE data with information gathered using alternative measurements. Most American JPE programs already make use of other information sources, such as narrative feedback, courtroom observation, reviews of the judge's written orders and opinions for clarity, and workload and other related caseload statistics. In a strong multi-method JPE program, alternative measurement methods used to complement JPE data should assess aspects of judicial performance that are not effectively captured in survey form. The shortcomings of each individual approach to evaluation should be balanced by strengths in other information sources used. Depending on measurement design and implementation practices, however, many common alternative measurement methods may also suffer from systematic gender and racial bias. Unless the potential for systematic bias is carefully considered in each new methodological approach used to inform an overall appraisal of judicial performance, the problem could be amplified as multiple sources of similarly biased information may reinforce and perpetuate discriminatory conclusions about actual judicial performance. In this section, we discuss the potential for bias in two other commonly used qualitative approaches to evaluating judicial performance: narrative feedback and courtroom observation.

Narrative feedback. A common qualitative method for gathering performance-relevant information in the United States is through the use of narrative feedback. Instead of asking respondents to rate the judge, an open-ended question is posited to solicit written comments about the judge's performance that the judge may receive verbatim for review. Less frequently, written comments are reportedly summarized to the judge or interpreted through feedback from a council of reviewers. In some states, commission members may review judges up for retention by soliciting citizen feedback through confidential written submissions or in person at public hearings. Recent performance appraisal studies have shown some systematic differences in the content of written feedback about women and minority group employees compared with their male majority counterparts (Biernat *et al*. 2012, Wilson 2010). In one study, supervisors emphasized different types of performance behaviors such as interpersonal or social skills of ethnic minority employees rather than their technical competence (Wilson 2010). In another, supervisors commented with greater frequency about the likelihood for promotion for male attorneys compared to their female counterparts, illustrating male attorneys' increased chances of becoming a partner in a Wall Street law firm

(Biernat *et al*. 2012). Although additional research should examine these issues in the context of written and verbal judicial feedback from court users, existing research suggests that improvements to the structure and criteria used for evaluation in traditional narrative evaluation formats may be needed to ensure fairness in the appraisal process, with implications for both professional development (by focusing on different skill sets) and future advancement.

Courtroom observation. Traditionally used in the United States by "good government" groups who package their comments into voters' pamphlets or media stories to influence retention elections (McCoy and Jahic 2006), courtroom observation is another methodology currently employed in some official state-sponsored JPE programs. As part of this process, trained laypersons may observe judicial performance in the courthouse and complete evaluation forms designed to capture the observer's assessment of the judge's performance that day, or judicial performance may be videotaped for subsequent review. The degree to which courtroom observation is conducted in a systematic manner varies across these programs. Among retention election states, Colorado and Utah have been leaders in developing observation-based data to incorporate into the evaluation process. The Utah observation program relies upon volunteers who are trained to rate the extent to which the judge was observed in court to comply with criteria drawn from procedural justice theory (see Woolf and Yim 2011). However, the criteria used in a courtroom observation program may require observers to rate judges in a manner that, like the survey-based measures previously discussed, involve inferences about several generic qualities of the judge. This may produce systematically biased results if criteria for evaluation are too abstract or vague, in the absence of a comprehensive, rigorous, ongoing training program for courtroom observers, or in absence of other quality assurance strategies. Depending on how courtroom observers are trained and how the reporting process is structured, their perceptions and judgments about judicial performance behavior may be influenced by implicit stereotypes and, particularly in the legal field, gendered performance expectations. Researchers should carefully assess courtroom observation programs and explore modifications to address the potential for systematic bias.

A note on other measurement approaches. Although we examine three qualitative measures of judicial performance in this paper, we note that quantitative measures, commonly but erroneously assumed to be "objective" measures, may also produce results that reflect and reinforce systematic biases. [9] Because these quantitative measures "should not be uniform across courts and judges" (Posner 2005, p. 1275) and require careful theoretical development to take into consideration the nuances of the broader context in which these measures are used, a comprehensive discussion of this complex issue falls beyond the scope of this paper. Economists attempting to devise these proxy measures, however, readily acknowledge that quantitative JPE measures may inadvertently discriminate in favor of or against certain types of judges based on qualities external to normative criteria associated with good judging (e.g., Smyth 2005, Choi and Gulati 2008, Choi *et al*. 2011). These measures – and the context in which they are used – must be carefully examined to ensure that the resulting evaluation is a fair one.

Of course, a rigorous examination of the fairness of a JPE program requires that those responsible for evaluating judges reveal their processes for doing so. This is not always the case. For example, the ABA Standing Committee conducts their own review of nominees for federal judgeships in the United States, but is not transparent about their assessment process. An analysis of their ratings of

[9] Many types of quantitative measures have been suggested, used, and hotly debated in efforts to better tap into various aspects of judicial performance, including judicial workload statistics, number of published judicial opinions, citation frequency, and reversal rates. These measures can sometimes result in an overemphasis on productivity in overall assessments of judicial performance. For example, see also the paper by Contini *et al*. (2014) in this issue.

Oñati Socio-legal Series, v. 4, n. 5 (2014), 863-879
ISSN: 2079-5971

nominees over the last 40 years reveals a systematic bias against women and minority candidates (Sen 2012). Importantly, these ABA ratings can influence the composition of the federal bench: The study also found a significant relationship between ABA ratings and the likelihood of Supreme Court appointment confirmation by the U.S. Senate. We call for the complete transparency of any judicial evaluation process in order to allow evaluation methodologies to be examined for systematic bias and, if necessary, improved.

6. Conclusion

The potential problem of stereotypic bias, although best understood in survey-based measures of judicial performance, is applicable to a range of available performance evaluation methodologies. This potential requires careful monitoring of patterns in performance evaluation data. It also requires the flexibility to respond to evidence of systematic bias through a careful reexamination and revision of the JPE program as a whole.

The use of biased JPE data may have a significant, pervasive, undesirable impact on the composition and quality of the judiciary and public perceptions of its legitimacy. When JPE results are used to guide decision-making by retention committees, the voting public, and others, biased data indicating that certain judges are better-suited or ill-suited for the position may disproportionately and unfairly jeopardize the careers of minority and female judges, thereby reducing the diversity of the bench (cf. Biernat *et al*. 2012, Sen 2012). Upon seeing significant racial and gender disparities in the composition of the judicial authority, the public may lose trust in a system that cannot appear to uphold its own ideals of fairness and justice. Biased evaluation results may also erode the self-worth of disadvantaged judges over time, cultivating the ideal environment for a "self-fulfilling prophecy" by which stereotypic criticisms are no longer merely a product of evaluators' perceptions, but instead become realized in actual behavior (see Snyder *et al*. 1977). Thus biased evaluation data may potentially damage the quality of judges on the bench in states that use a JPE program to facilitate professional development.

Our perspective assumes that judicial performance quality does not consistently differ by gender, an assumption that is generally supported in the empirical literature (see Boyd *et al*. 2010, Choi *et al*. 2011) despite observed differences in survey ratings (e.g., Gill *et al*. 2011 Sen 2012). The work reviewed in this paper suggests that the impact of bias on evaluations of performance may be lessened by focusing evaluators' attention on a judge's observed behavior. This emphasis on the rater's direct professional working experience with a judge limits the role of implicit bias in evaluations. A focus on observable behavior also minimizes the role in these types of measures for the assessment of "appropriate" personality traits or temperament, criteria that are quite popular in the United States and elsewhere, yet which in our view are wholly inappropriate for evaluating the on-the-job performance of judges. This makes our approach to improving judicial performance evaluation surveys difficult to reconcile with trait-based perspectives. It also may not be wholly compatible with the viewpoint of those interested in diversifying the bench on the premise that women and minority judges contribute unique perspectives or nontraditional judging styles that result in better quality judicial decisions. For example, after finding evidence of gender bias in his work to apply the ABA model to Australian judges, Colbran (2002, p. 68) concluded that "women judges should be evaluated differently from male counterparts, with criteria and measures sensitive to gender issues." The presumption is that female judges have different decision-making processes or styles that a specialized assessment, designed to alter the standards for evaluating female judges and better tap into "gender issues," would capture. In our view, introducing such a segregated approach to JPE could in itself serve to legitimize some forms of stereotypic bias that could ultimately prove harmful to the status of women in the judiciary.

As programs for the evaluation of judicial performance grow and spread, efforts to improve measurement should be matched with equally rigorous conceptual development. Popular views of "good judging" in the United States have been primarily defined by the ABA model, which despite a paucity of empirical validation continues to influence the content of JPEs nationally and internationally. Researchers and practitioners should continue to consider and investigate other conceptualizations of good judging in concert with the kinds of efforts outlined in this paper.

References

Baltes, B.B., Bauer, C.B., and Frensch, P., 2007. Does a structured free recall intervention reduce the effect of stereotypes on performance ratings and by what cognitive mechanism? *Journal of Applied Psychology*, 92 (1), 151-164.

Bauer, C.C. and Baltes, B.B., 2002. Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles*, 47 (9-10), 465-476.

Biernat, M., Tocci, M.J., and Williams, J.C., 2012. The language of performance evaluations: Gender-based shifts in content and consistency of judgment. *Social Psychological and Personality Science*, 3 (2), 186-192. doi:10.1177/1948550611415693

Biernat, M., and Kobrynowicz, D., 1997. Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, 72 (3), 544–557.

Biernat, M., Vescio, T.K., and Manis, M., 1998. Judging and behaving toward members of stereotyped groups: A shifting standards perspective. *In*: C. Sedikides, J. Schopler, and C.A. Insko, eds. *Intergroup cognition and intergroup behavior*. Hillsdale, NJ: Lawrence Erlbaum Associates, 151-175.

Boyd, C.L., Epstein, L., and Martin, A.D., 2010. Untangling the causal effects of sex on judging. *American Journal of Political Science* [online], 54 (2), 389–411. doi: 10.1111/j.1540-5907.2010.00437.x. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2010.00437.x/pdf [Accessed 17 December 2014]

Brody, D.C., 2008. The use of judicial performance evaluation to enhance judicial accountability, judicial independence, and public trust. *Denver University Law Review* [online], 86 (1), 115-156. Available from: http://www.law.du.edu/documents/denver-university-law-review/v86_i1_brody.pdf [Accessed 17 December 2014].

Burger, G.K., 2007. *Attorney's ratings of judges: 1998-2006*. Report to the Mound City Bar. Mound City, MO.

Casey, P., *et al*., 2012. *Helping courts address implicit bias: Resources for education* [online]. Williamsburg, VA: National Center for State Courts. Available from: http://www.ncsc.org/IBReport [Accessed 17 December 2014].

Cermak, K. and Block, R., 2001. Perceptions of good and bad judging: An analysis of the Illinois Judicial Development Project. *In:* J. Van Hoy, ed. *Legal Professions: Work, Structure, and Organization*. London: Elsevier Press, 253-270.

Choi, S.J., *et al*., 2011. Judging women. *Journal of Empirical Legal Studies*, 8 (3), 504–532. doi: 10.1111/j.1740-1461.2011.01218.x.

Choi, S.J. and Gulati, M., 2008. Bias in judicial citations: A window into the behavior of judges? *The Journal of Legal Studies*, 37 (1), 87-130.

Clark, L.A. and Watson, D., 1995. Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7 (3), 309-319.

Colbran, S., 2002. Queensland magistrates' judicial development project. *In*: *2nd Annual AIJA Magistrates' Conference, 13-14 September 2002* [online]. North Quay, Brisbane, Australia. Available from: http://www.aija.org.au/Mag02/Stephen%20Colbran.pdf [Accessed 17 December 2014].

Contini, F., Mohr, R., Velicogna, M., 2014. Formula over Function? From Algorithms to Values in Judicial Evaluation. *Oñati Socio-legal Series* [online], 4 (5), 1099-1116. Available from: http://ssrn.com/abstract=2533902 [Accessed 23 December 2014].

Deaux, K. and Emswiller, T., 1974. Explanations of successful performance in sex-linked tasks: What is skill for the male is luck for the female. *Journal of Personality and Social Psychology*, 29 (1), 80-85.

Dillman, D.A., Smyth, J.D., and Christian, L.M., 2009. *Internet, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: Wiley.

Dunning, D., and Sherman, D.A., 1997. Stereotypes and tacit inference. *Journal of Personality and Social Psychology*, 73 (3), 459–471.

Durham, C.M., 2000. Gender and professional identity: Unexplored issues in judicial performance evaluation. *Judges' Journal*, 39 (2), 13-16.

Eagly, A.H., Wood, W., and Diekman, A.B., 2000. Social role theory of sex differences and similarities: A current appraisal. *In*: T. Eckes and H.M. Trautner, eds. *The developmental social psychology of gender*. Mahwah, NJ: Erlbaum, 123–174.

Eagly, A.H. and Karau, S.J., 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109 (3), 573-598.

Elek, J.K., Rottman, D.B., and Cutler, B.L., 2012. Judicial performance evaluation: Steps to improve survey process and management. *Judicature* [online], 96 (2), 65-75. Available from: http://www.ncsc.org/~/media/Files/PDF/Information%20and%20Resources/65-75_Elek_962.ashx [Accessed 17 December 2014].

Elek, J.K., and Rottman, D.B., 2013. Improving Judicial-Performance Evaluation: Countering Bias and Exploring New Methods. *Court Review*, 49 (3), 140-144.

Englich, B., Mussweiler, T., and Strack, F., 2006. Playing dice with criminal sentences: the influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, 32 (2), 188–200.

Flanders, S., 1978. Evaluating the judges: How should the bar do it? *Judicature*, 61 (7), 304-310.

Gill, R., 2012. *Judicial performance evaluations as biased and invalid measures: why the ABA Guidelines are not good enough* [online]. Available from: http://ssrn.com/abstract=2031800 [Accessed 23 October 2013].

Gill, R.D., Lazos, S.R., and Waters, M.M., 2011. Are judicial performance evaluations fair to women and minorities? A cautionary tale from Clark County, Nevada. *Law & Society Review*, 45 (3), 731–759. doi: 10.1111/j.1540-5893.2011.00449.x.

Gill, R.D. and J. Retzl, K.J., 2013. Implicit gender bias in state-sponsored judicial performance evaluations: A preliminary analysis of Colorado's JPE system, 2002-2012 [online]. *Paper presented at the Annual Meeting of the Western Political Science Association, Hollywood, CA*. Available from: http://ssrn.com/abstract=2270376 [Accessed 20 December 2014].

Greenwald, A.G., and Krieger, L.H., 2006. Implicit bias: Scientific foundations. *California Law Review* [online], 94 (4), 945–967. Available from: http://scholarship.law.berkeley.edu/californialawreview/vol94/iss4/1/ [Accessed 17 December 2014].

Guthrie, C., Rachlinski, J.J., and Wistrich, A.J., 2001. Inside the judicial mind. *Cornell Law Review* [online], 86 (4), 777-830. Available from: http://ssrn.com/abstract=257634 [Accessed 17 December 2014].

Guterman, J.H. and Meidinger, E.E., 1977. *In the opinion of the bar: A national survey of bar polling practices*. Chicago, IL: American Judicature Society.

Institute for the Advancement of the American Legal System, 2008. *The bench speaks on judicial performance evaluation: A survey of Colorado judges* [online]. Denver, CO: Institute for the Advancement of the American Legal System, University of Denver. Available from: http://iaals.du.edu/images/wygwam/documents/publications/Bench_Speaks_On_JPE2008.pdf [Accessed 20 December 2014].

Johnson, C., *et al*., 2007. Feeling injustice, expressing injustice: How gender and context matter. *Advances in Group Processes*, 24, 149-186.

Kearney, R.C. and Sellers, H., 1996. Sex on the docket: Reports of state task forces on gender bias. *Public Administration Review*, 56 (6), 587-592.

Knowlton, N. and Reddick, M., 2012. *Leveling the playing field: Gender, ethnicity, and judicial performance evaluation* [online]. Report for the Institute for the Advancement of the American Legal System. Denver, CO: Institute for the Advancement of the American Legal System, University of Denver. Available from: http://iaals.du.edu/images/wygwam/documents/publications/IAALS_Level_the_Playing_Field_FINAL.pdf [Accessed 17 December 2014].

Levinson, J.D., and Young, D., 2010. Implicit gender bias in the legal profession: An empirical study. *Duke Journal of Gender Law & Policy* [online], 18 (1), 1-44. Available from: http://scholarship.law.duke.edu/djglp/vol18/iss1/1 [Accessed 17 December 2014].

Levy, M.K., Stith, K., and Cabranes, J., 2010. The costs of judging judges by the numbers. *Yale Law & Policy Review* [online], 28 (2), 313-24. Available from: http://digitalcommons.law.yale.edu/fss_papers/1289 [Accessed 17 December 2014].

Malcolm, D., 1994. *Report of Chief Justice's task force on gender bias*. Perth: Supreme Court of Western Australia.

Martell, R.F., 1996. What mediates gender bias in work behavior ratings? *Sex Roles*, 35 (3-4), 153–169.

McCoy, C. and Jahic, G., 2006. Familiarity breeds respects: Organizing and studying a Courtwatch. *Justice System Journal*, 27 (1), 61-70.

Nunnaly, J., 1978. *Psychometric theory*. New York: McGraw-Hill.

Posner, R.A., 2005. Judicial behavior and performance: An economic approach. *Florida State University Law Review*, 32 (4), 1259-1279.

Rachlinski, J.J., *et al*., 2009. Does unconscious bias affect trial judges? *Notre Dame Law Review* [online], 84 (3), 1195-1246. Available from: http://ssrn.com/abstract=1374497 [Accessed 17 December 2014].

Resnik, J., 1996. Asking about gender in courts, *Signs: Journal of Women in Culture and* Society, 21 (4), 952-990.

Rudman, L. A. and Glick, P., 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57 (4), 743-762.

Schotland, R.A., 2006. A sporting proposition - SEC v. Chenery. *In:* P.L. Strauss, ed. *Administrative Law Stories*. New York, NY: Foundation Press, 168-188.

Scott, K.A. and Brown, D.J., 2006. Female first, leader second? Gender bias in the encoding of leadership behavior. *Organizational Behavior and Human Decision Processes*, 101 (2), 230-242.

Sen, M., 2012. Below the bar? Racial and gender bias in judicial nominations [online]. *Paper presented at the 2012 Visions in Methodology Conference, State College, PA*. Available from: http://visionsinmethodology.org/wp-content/uploads/2012/04/Sen_VIM2012.pdf [Accessed 23 October 2013].

Smyth, R., 2005. Do judges behave as *homo economicus* and, if so, can we measure their performance? An antipodean perspective on a tournament of judges. *Florida State University Law Review* [online], 32 (4), 1299-1330. Available from: http://www.law.fsu.edu/journals/lawreview/downloads/324/Smyth.pdf [Accessed 22 December 2014].

Snyder, M., Tanke, E.D., and Berscheid, E., 1977. Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, 35 (9), 656-666.

Sterling, J.S., 1993. The impact of gender bias on judging: Survey of attitudes toward women judges. *Colorado Lawyer*, 22 (February), 257-258.

Strickland, S., *et al*., 2012. *State Court Organization* [online]. Available from: http://data.ncsc.org/QvAJAXZfc/opendoc.htm?document=Public%20App/SCO.qvw&host=QVS@qlikviewisa&anonymous=true&bookmark=Document\BM21 [Accessed 23 October 2013].

Tourangeau, R., 1984. Cognitive science and survey methods. *In*: T. Jabine et al., eds. *Cognitive aspects of survey methodology: Building a bridge between disciplines*, Washington, DC: National Academy Press, 73-100.

Tyler, T.R., 1990. *Why people obey the law*. New Haven, CT: Yale University Press.

Tyler, T.R., 2007. *Legitimacy and Criminal Justice*. New York: Russell Sage Foundation.

Uhlmann, E., and Cohen, G. L., 2005. Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, 16 (6), 474-480.

Wilson, K., 2010. An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations*, 63 (12), 1903-1933.

Wistrich, A.J., Guthrie, C., and Rachlinski, J.J., 2005. Can judges ignore inadmissible information: The difficulty of deliberately disregarding. *University of Pennsylvania Law Review*, 153 (4), 1251-1346.

Woolf, N., and Yim, J., 2011. The courtroom-observation program of the Utah Judicial Performance Evaluation Commission. *Court Review* [online], 47 (4), 84-91. Available from: http://digitalcommons.unl.edu/ajacourtreview/368/ [Accessed 17 December 2014].