2014

Implicit Bias in Judicial Performance Evaluations: We Must Do Better Than This

Rebecca D. Gill
University of Nevada, Las Vegas, rebecca.gill@unlv.edu

Justice System Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ujsj20

Implicit Bias in Judicial Performance Evaluations: We Must Do Better Than This

Rebecca D. Gill[a]
[a] Department of Political ScienceUniversity of Nevada Las Vegas, Las Vegas, Nevada
Published online: 09 May 2014.

To cite this article: Rebecca D. Gill (2014): Implicit Bias in Judicial Performance Evaluations: We Must Do Better Than This, Justice System Journal, DOI: 10.1080/0098261X.2013.873290

To link to this article: http://dx.doi.org/10.1080/0098261X.2013.873290

JUSTICE SYSTEM JOURNAL, *00*(0), 1–24, 2014

ISSN: 0098-261X print / 2327-7556 online
DOI: 10.1080/0098261X.2013.873290

Routledge
Taylor & Francis Group

Implicit Bias in Judicial Performance Evaluations: We Must Do Better Than This

Rebecca D. Gill

Department of Political Science, University of Nevada Las Vegas, Las Vegas, Nevada

Judicial performance evaluations (JPEs) are a critical part of selecting judges, especially in states using merit-based selection systems. This article shows empirical evidence that gender and race bias still exist in attorney surveys conducted in accordance with the ABA's Guidelines. This systematic bias is related to a more general problem with the design and implementation of JPE surveys, which results in predictable problems with the reliability and validity of the information obtained through these survey instruments. This analysis raises questions about the validity and reliability of the JPE. This is a particularly poor outcome, as it means that we are subjecting many judges to state-sponsored evaluations that are systematically biased against women and minorities.

KEYWORDS: judicial performance evaluation, implicit bias, judicial selection, judicial quality, gender bias, race bias

INTRODUCTION

Judicial performance evaluations (JPEs) are a critical part of selecting judges, especially in states using merit-based selection systems. These JPEs, which have been increasing in popularity in recent years, fulfill two important functions. First, JPEs provide voters with information upon which they rely when casting their ballot in judicial elections (Bernick and Pratto 1995; Esterling 1999; Hall 1985). Second, the evaluation of judicial performance is an important tool for protecting the quality and accountability of judges (White 2009). Ideally, JPE programs should be "designed and administered in a way that does not inadvertently harm the principles they are intended to promote" (Esterling 1999, 207), especially in light of the efforts of various "good government" interest groups to implement merit-based selection systems in states around the country (Cann 2006).

In an effort to help states pursue these goals, the American Bar Association published its "Guidelines for the Evaluation of Judicial Performance" (American Bar Association 1985, 2005). All of the current and model JPE systems have been informed by these Guidelines, which admonish states to devise JPEs that supply judges and voters with reliable and unbiased information. A number of "behavior-based measures" of judicial performance for use in attorney surveys have been derived from the Guidelines. Such measures feature prominently in other model JPE systems

Address correspondence to Rebecca D. Gill. Assistant Professor, Department of Political Science, University of Nevada Las Vegas, 4505 S. Maryland Parkway, Box 455029, Las Vegas, NV 89154-5029. E-mail: rebecca.gill@unlv.edu

(IAALS 2006a, 2006b; Puiszis 2011), as well as in all of the state-sponsored (Gill et al. 2011) and many of the bar association-sponsored attorney surveys (e.g., Brody 2012, 2008a; Brody and Lovrich 2009). But the behavior-based evaluation measures derived from the language of the Guidelines have not been validated through rigorous analysis. A recent scholarly exposition of the methods used in JPE surveys concludes that "these surveys tended to share some fundamental design flaws that could produce low-quality data" (Elek et al. 2012, 67). Although these survey design problems can yield all manner of troublesome biases and inconsistencies in the resulting data, none is starker than the systematic discrepancies in survey results between judges based on their race and gender.

Research is only now beginning to address the critical question of how to minimize gender and race bias in the judicial evaluation process (Durham 2000). Of the twenty-two states that have state-sponsored JPE systems in place, just one has undertaken the work of reviewing JPE surveys for empirical evidence of systematic gender and race bias, and that study is almost twenty years old.[1] The present analysis shows empirical evidence that gender and race bias still exist in attorney surveys conducted in accordance with the ABA's Guidelines. This systematic bias is related to a more general problem with the design and implementation of JPE surveys, which results in predictable problems with the reliability and validity of the information obtained through these survey instruments.

JUDICIAL PERFORMANCE EVALUATION

Although evaluations of judicial performance are nothing new (Feeney 1987), there are increasing calls for reliance on JPEs. Proponents of JPE cite a number of important potential benefits. Scholars have long recognized the voter information problem in judicial elections (Dubois 1980). The problem is especially pronounced in retention elections (Hojnacki and Baum 1992). Voters in judicial elections report a desire for more information about judicial candidates (Boatright and Esterling 2000). JPEs are seen as an important way to solve the voter information problem in judicial elections (Hall and Aspin 1987). They accomplish this by providing voters with relevant information about the performance of their sitting judges (Brody 2008b), even in states with competitive judicial elections (Brody 2003).

Scholars also argue that JPEs can protect the quality of the judges on the bench, in part by encouraging self-improvement (Aynes 1981; Chauvin 1989). Indeed, many judges report that the performance evaluation process helps them become better judges (Esterling 1999). This individual self-improvement can result in aggregate level improvement in the administration of justice from the entire judiciary (Feeney 1987). Even vocal critics of JPE agree that the use of such tools confidentially for the purposes of self-improvement can be a great benefit to judges (Griffin 1994).

By subjecting judges to performance evaluations, states can reinforce the idea that judges are not like other politicians, and this may help to increase the legitimacy of the courts more generally (White 2009). The process can help "to remind citizens of the judges' unique role in democratic

[1]A 1993 study of the results of the Colorado Judicial Performance Evaluation Commission's lawyer survey showed that male and female lawyers alike rated female judges consistently lower than male judges (Sterling 1993). Colorado has since adjusted its evaluation methods, but no rigorous follow-up studies have been conducted to confirm that the disparities have been resolved.

government" (Esterling 1999, 215). JPEs can encourage voters to think of judicial quality in terms of the criteria outlined in the ABA Guidelines instead of evaluating judges through the prism of policy-based outcomes (Kourlis and Singer 2007; Singer 2007). Some scholars argue that, when given the appropriate information, voters may be inclined to view concerns about procedural justice as relevant to their evaluation of judicial candidates (Olson 2001).

Underlying all of these justifications for JPE is the assumption that the evaluation process can yield important determinations about the quality of the performance of individual judges on the bench. Most observers of the courts know a high- or low-quality judge when they see one, but it is a much more difficult task to design a rigorous instrument to define and measure judicial quality systematically. In its 1985 publication, "Guidelines for the Evaluation of Judicial Performance" (American Bar Association 1985; updated in 2005, collectively "the Guidelines"), the ABA laid out a series of criteria by which observers should measure the quality of judges. These criteria are divided into a series of categories, including the judge's legal ability, integrity and impartiality, communication skills, professionalism and temperament, and administrative capacity. These categories of judicial quality are not controversial, and they are used widely as the basis of judicial performance evaluation programs throughout the United States (Gill et al. 2011). They represent the goals we have for judicial officers operating at peak performance levels.

The Guidelines provide an important starting place, but they give little guidance when it comes to creating operational measures of any of these broad dimensions of performance. The 2005 commentary recommends very strongly that the design of any specific measurement instruments be left to experts, and that behavior-based instruments be used to "generate more meaningful information about judicial behavior" (American Bar Association 2005, 13). Although most state-sponsored JPEs feature a number of different sources of information, the survey instrument is "clearly the key component of all JPE programs" (Brody 2000, 336). Commission members report relying heavily on the attorney surveys in the production of commission recommendations (Esterling and Sampson 1998). This is, in part, because "technical competence is a primary criterion and cannot be well assessed by those without that technical competence themselves" (Olson and Batjer 1999, 129).

RELIABILITY AND VALIDITY IN SURVEY-BASED PERFORMANCE ASSESSMENT

It is difficult to obtain accurate information about job performance. Performance rating systems "almost inevitably contain various forms of rater errors" (Borman 1978, 135). Although research on performance appraisal has flourished, "reliability and validity still remain major problems in most appraisal systems" (Banks and Murphy 1985, 335). JPEs come with a unique combination of reliability and validity issues because of their unique survey structure.

Across all sectors of society in which performance appraisal is conducted, the process is generally disliked by everyone involved. Indeed, it "seems to be tolerated only because no one can think of any realistic, better alternatives" (Shafritz et al. 1986, 428). The process of performance appraisal in the public sector is different from its private sector cousin (Daley 1992). It is particularly important to be diligent about performance assessment design in the public sector because "the business of choosing [performance] measures is the business of choosing what government does" (Moynihan et al. 2011, 152). When we do performance appraisal badly,

"we end up playing God with an employee's career on the basis of little more than a personal opinion" (Morrisey 1983, 4).

The literature on performance appraisal is diverse, spreading over a number of academic disciplines (Daley 1992). Much of the research on improving the reliability of performance appraisals concludes that evaluator training programs can help to minimize measurement error problems (Bernardin et al. 2012; Daley 1992; Landy and Farr 1980; Martin and Bartol 1988; Morrisey 1983). About 90 percent of job performance appraisals of public sector workers are conducted by supervisors (Daley 1992).[2] But JPEs differ from most performance appraisals in that they are not conducted by the judge's supervisor. The attorney survey evaluation of judicial performance most closely resembles something between a peer review and a review by subordinates. The lawyers and judges are not exactly peers in the workgroup setting, but they do share in common an educational background and a practical understanding of what it takes to be a good judge.

In light of the practical impossibility of providing a broad-based rater training program for respondent attorneys, it is particularly important for JPE designers to be attentive to the potential for measurement error. Errors in measurement can yield data that do not reflect the underlying reality that the performance appraisal tool is purporting to measure. Measurement error comes from the interaction of deficiencies in the instrument and characteristics of the respondent (Bautista 2010). For example, response errors such as leniency error, the halo or horns effect, and central tendency bias can all result from instruments that insufficiently define the performance standards or response options (Bernick and Pratto 1995). The type of error that results often depends on the particular characteristics of the evaluator (Landy and Farr 1980).

Most JPE surveys use something called a graphic rating scale, which gives a Likert-style set of response items for a set of (mostly) behavioral patterns. JPE surveys typically use adjectival or numerical anchors to guide raters in choosing the appropriate response. These anchors may be arranged in the style of grades from A to F, as in Colorado's survey. Alternatively, they can be broad descriptions of acceptability, as in a response set of "not adequate," "adequate," and "more than adequate." In each case, the meaning of these anchors is vague, and their interpretation will vary across raters (Daley 1992).

The ABA Guidelines (American Bar Association 2005) recommend transforming these graphic rating scales into "behavior-based measures." Although the Guidelines do not mention it by name, the description of these measures is much like the behaviorally anchored rating scales (BARS) procedure.[3] Using a critical incident technique, BARS instrument designers conduct a thorough job analysis and identify concrete examples of job-related behaviors. Then these behaviors are set to scale using descriptive anchors that represent a continuum of possible levels of performance.

The job analysis feature of the BARS development procedure is an important bulwark against many validity problems. Content validity is boosted through this process, which casts a wide net for examples of important job-related behavior. The job analysis also includes an expert determination via consensus about the weights that each individual behavior dimension should

[2]JPEs are one of the few instances where surveys are used for formal job performance evaluation. As such, the sampling error problems of survey research complicate the already difficult measurement error problems inherent in employee performance appraisal. It is beyond the scope of this article to address the survey-related problems with unrepresentative data. For a discussion of these issues, see Elek et al. (2012) and Wood and Lazos (2009).

[3]Since the creation of the BARS procedure (Smith and Kendall 1963), a number of similar systems have been developed that share a number of the same basic features. For a brief description, see Prowse and Prowse (2009). For the purposes of this article, the term BARS will be inclusive of these progeny.

be accorded in an overall assessment of job performance. Construct validity can suffer when the performance dimensions measured by the instrument cannot be distinguished by the evaluators. The job analysis allows for a team of experts to identify and isolate the important constructs that underlie expectations of the various levels of performance excellence for the job. These improvements in content and construct validity should reveal themselves in an increased criterion validity, whereby external, objective measures of performance should be related to the measures on the performance appraisal instrument.

The BARS approach can also mitigate problems with measurement error. Response biases such as leniency and central tendency error are reduced by removing from the raters the responsibility for defining the adjectival anchors for themselves (Bernick and Pratto 1995). The BARS strategy counteracts the halo/horns effects by defining the various dimensions precisely (Bernick and Pratto 1995). Interrater reliability is also improved, as differences in rater definitions of the adjectival or numeric anchors are neutralized (Bernick and Pratto 1995). In short, this strategy is thought to improve both the validity and the reliability of the instrument.

To at least some degree, the development of the BARS technique was spurred on by the decisions of judges in employment discrimination cases (Stryker et al. 2011). Courts, convinced by the testimony of plaintiffs' expert witnesses, have pushed for the replacement of the more subjective performance evaluations with objective appraisals based on behaviors or management objectives. This is because, along with the other purported benefits of BARS systems, they are also thought to decrease the incidence of explicit and implicit bias on the basis of race, gender, or other immutable characteristics.

Although the ABA Guidelines (American Bar Association 2005) recommend the use of behaviorally anchored response scales, this suggestion has not been followed in practice. The JPE attorney surveys that have been derived from the ABA Guidelines almost universally purport to use "behavior-based measures" to assess judicial performance (Brody 2008a, 2012; Brody and Lovrich 2009; IAALS 2006a, 2006b). These measures, however, are not BARS; they lack the particular job analysis and behavioral anchoring that are at the core of the BARS evaluation instrument (Daley 1992).

A recent analysis of eighteen currently used JPEs and four model JPE systems shows that the survey instruments suffer from widespread problems; many of these instruments confound multiple aspects of judicial behavior, ask questions using imprecise language, and provide little helpful anchoring information in the response categories (Elek et al. 2012). Each of these failures can contribute to measurement error by forcing evaluators to make judgment calls about the meaning of the question prompt, the kind of evidence that supports a high or low rating on that question prompt, and the appropriate response category that represents a particular rating.

In this situation, the instrument fails to give the evaluator enough information to make a judgment. The evaluator supplements the instrument using information that is less relevant—and perhaps inappropriate—to the task at hand. This is how respondent characteristics work together with the inadequate instrument to create measurement error. And, in many instruments, the measurement error reflects the implicit biases of the evaluator.

IMPLICIT BIAS IN JPE

Traditionally, gender and race discrimination have been understood as products of conscious motive or intent (Krieger 1995). More recently, scholars have recognized the important impact

of unconscious or implicit bias on the perception of performance in gender and race-stereotyped jobs (Gill et al. 2011). Male legal professionals tend to perceive much less gender bias in the workplace than do their female colleagues (Coontz 1995). Even in the context of the popularity of diversity initiatives in law schools, race-based stereotypes of law students have a disproportionally negative effect on minority students (Clydesdale 2004). Indeed, achievement levels for minority lawyers still lag, even in the face of economic incentives for law firms to increase racial diversity (Gordon 2003).

Social science research, especially in the field of cognitive psychology, has identified a more innocent but pernicious cause of gender and race discrimination: implicit bias. The process of simplifying and categorizing our environment, which exists as a necessary condition for most higher-level cognitive function, processes people just as it does letters, shapes, and colors (Lee 2005). Even absent a conscious bias against women or minorities, everyone is exposed to the societal stereotypes associated with different categories of people. It is through the lens of these stereotypes that we perceive, process, store, recall, and synthesize information about people. Our actions may be based, in part, on the accumulated stereotypes about a particular outgroup, resulting in inaccuracy and unfairness based on race or gender.

The social science evidence for implicit race and gender bias in employment decisions is strong and convincing. In fact, this theory of decision making played a pivotal role in the Supreme Court's decision in *Price Waterhouse v. Hopkins* (1989) [490 U.S. 228], which held that gender stereotypes had been used to deny a female accountant's bid for partner (Fiske et al. 1991). Social cognition theory explains that humans are naturally programmed to apply cognitive schemas to aspects of our interpersonal relationships. Just as we use situational stereotypes as shortcuts to understanding our physical world, we also develop them to organize our interpersonal interactions.

This works nicely when we are aware of what we are doing and when we can control the content and activation of these schemas. But implicit social cognition theory holds that this is not usually the case; instead, we are gathering information and categorizing people at a subconscious or unconscious level. Implicit cognition is "the process through which we become sensitive to certain regularities in the environment (1) in the absence of intention to learn about those regularities, (2) in the absence of the awareness that one is learning, and (3) in such a way that the resulting knowledge is difficult to express" (Cleeremans 2003, 491). Implicit social cognition is the application of this cognitive process to information about groups of people.

This is what gives rise to implicit bias. This kind of bias happens much more furtively than bias based on explicit racism or sexism. People who self-report low levels of racial or gender bias can still exhibit implicit bias driven by underlying stereotype schemas (Lee 2005). This does not mean that self-reported measures of sexism and racism are disingenuous; instead, people are "unable to know the contents of their mind" (Kang and Banaji 2006, 1071), and the stereotypes creep in to frame evaluations of others without our conscious consent.

A few aspects of implicit social cognition theory are particularly relevant to JPEs. First, higher rates of bias tend to occur in hiring-related decisions where the characteristics that are stereotypical for the job are at odds with the gender or race stereotype (Heilman 1983; Landy and Farr 1980). This kind of stereotype can survive repeated exposure to individual judges whose behavior and performance does not seem to conform to the underlying race or gender stereotypes of the profession (Carnes et al. 2005). This often results in a paradox or "double bind" for women in the legal profession because they continue to be penalized in their performance evaluations

both for being too masculine and for not fitting the masculine stereotype of the job (Bowman 1998).

A second important characteristic of implicit bias is the fact that subjective, vague, or abstract evaluation criteria exacerbate discriminatory employment decisions (Fiske et al. 1991). In other words, more traditional reliability and validity deficiencies in performance appraisal instrument design leaves more room for implicit biases to drive the results. In JPEs, "[t]he force of traditional stereotypes is compounded by the subjectivity of performance evaluations" (Rhode 2001, 15). Previous research finds that the yes-or-no question, "Should Judge X be retained?" can have this effect (Gill et al. 2011). The work of judges and other legal professionals is often based at least partially on subjective assessments, "relying on the judgments of supervisors and colleagues regarding the less measurable activities" (Choi et al. 2009, 1319).

Other characteristics of the evaluation environment can exacerbate implicit race and gender bias. Anonymous performance ratings tend to have more variance (Landy and Farr 1980). Evidence suggests that the anonymity of evaluations increases the effects of implicit bias (Hekman et al. 2010), such that implicit bias may be the source of the larger variance. Evaluations that are done quickly are also more subject to this kind of bias (Carnes et al. 2005). Evaluations of performance after the fact can also encourage bias, as the evaluator is required to access stored memories in order to make an assessment. Information that is inconsistent with existing unconscious stereotypes is more difficult for the brain to store, but supporting evidence may be magnified in the memory, and even embellished or fabricated unknowingly (Bartlett 1932).

All of these conditions hold in attorney surveys of judicial performance. Judging, like the practice of law more generally, is a male-stereotyped activity. The types of questions asked are generally behavioral in nature, but the instruments often suffer from design defects that make it difficult for respondents to formulate objective assessments. These surveys are often done quickly, as attorneys are asked to rate several judges at a time on their performance over the past two years. The resulting assessments are anonymous. In all, surveys of judicial performance may be even more likely than other performance evaluations to suffer from unconscious gender and race bias.

The Guidelines (American Bar Association 2005) suggest the use of behavior-based measures in part to combat the effects measurement error in judicial performance evaluations. The Guidelines also note that, in correcting these measurement error problems, implicit bias in judicial performance evaluations will be mitigated. The Guidelines predict that using behavior-based measures will "reduce subjectivity in assessments of judicial performance, thus limiting the potential for gender and other biases to influence responses" (American Bar Association 2005, 14). But the JPE survey instruments derived from the Guidelines have failed to incorporate some of the most important features of the BARS procedure (Elek et al. 2012; Bernick and Pratto 1995), and there is reason to suspect that these surveys have not achieved the error and bias reductions that the ABA envisioned.

AN EMPIRICAL ANALYSIS OF JPE ATTORNEY SURVEY DATA

The problems of measurement error outlined above are systematically linked to implicit bias. Where the survey instrument provides too little guidance for respondents, measurement error is the result. Part of this measurement error is the gender and race bias that results from the

unconscious activation of stereotypes anticipated by implicit bias theory. On its face, the typical JPE survey instrument has the markers for the kind of measurement error that comes from poorly designed performance measures (Elek et al. 2012). In addition, there is evidence that the job of the judge carries with it a strong and pervasive set of race and gender stereotypes. Indeed, a slew of committees and task forces "have found that bias was prevalent in the experiences of female and minority litigants, court employees, and attorneys" (Knowlton and Reddick 2012, 28).

The presence of race and gender bias in JPE attorney survey results, then, is an indicator of broader problems with the performance evaluation instrument. Bias is likely in performance evaluations that lack "an easily accessible check list of objective cues for the evaluation of performance" (Wheery and Bartlett 1982, 534). Given the particular potential for implicit race and gender bias in the legal profession, the bias associated with poorly designed appraisal measures is likely to manifest, at least in part, as gender and race bias.

To get a clear sense of the extent of the problem, then, a multipronged analytical strategy is required. Ideally, individual attorney responses on performance evaluation questionnaires, accompanied by important demographic and other contextual information about the attorney respondents, would be analyzed. This type of data would allow for an investigation of a number of indicators of validity, reliability, and the nature of any race or gender gap in the resulting performance evaluation ratings.

At present, this is not practical. Although the Guidelines (American Bar Association 2005) and other model codes (IAALS 2006a, 2006b) stress the importance of wide dissemination of survey results, the recommended dissemination format includes only performance data aggregated by judge. This, of course, is out of concern for "ensur[ing] the anonymity of respondents to performance questionnaires" (American Bar Association 2005, 14). Some JPE commissions provide summary demographic information about the respondents, but none provides access to even redacted versions of the respondent-level data.

Even still, some important information relevant to the issues of implicit bias can be gained from close analysis of aggregate JPE survey data. Evidence of persistent differences in scores on the basis of race or sex that are not mitigated by the inclusion of objective information about judicial performance is consistent with the hypothesis that implicit bias is present in the survey results. It is likely that certain questions in the evaluation process trigger different gendered and raced understandings of what it means to perform that trait well. There is evidence that judicial temperament and legal knowledge survey questions introduce systematic implicit gender bias (Durham 2000).

Evidence that implicit bias occurs in tandem with other problems with validity and reliability adds credence to the idea that the gender and race bias is a symptom of broader problems with the JPE attorney surveys as currently implemented. Although it is not possible to identify the exact cause of any reliability or validity issues in aggregate data, many of these problems will manifest themselves in particular patterns at the aggregate level. For example, high average scores across all performance dimensions suggest a problem with the halo effect. A high correlation among aggregate scores on all of the different performance dimensions suggests an inability of the dimensions to isolate various important performance factors. If the interdimensional correlations are not significantly lower than the intradimensional correlations, this also suggests a problem with construct validity.

The attorney survey data in this analysis come from the *Las Vegas Review-Journal's* "Judging the Judges" survey. This survey of practicing attorneys is conducted every even year in conjunction

with the judicial election cycle. Although this particular survey is not part of a comprehensive state- or bar-sponsored JPE program, the survey was created by joint committee of experts (Hopkins 2012) and is very similar in design and implementation to existing state-sponsored attorney surveys (Gill et al. 2011).

The survey asks respondent attorneys to rate the performance of all the judges on the Clark County ballot before whom they have argued a case in the evaluation period. The resulting data are aggregated by judge for each year, yielding 350 average scores across 94 judges.[4] This is an unbalanced panel dataset, as not all judges served on the bench across the entire twelve-year survey period. The survey uses a set of twelve questions[5] to assess the five main performance dimensions suggested in Section V of the ABA Guidelines (see Table 1). These categories are Legal Ability (LA), Integrity and Impartiality (II), Communication Skills (CS), Professionalism and Temperament (PT), and Administrative Capacity (AC).

From the start, many of the problems identified by Elek, Rottman, and Cutler (2012) are present in the Judging the Judges survey. Take, for example, Question 11 from the 2002–2008 version of the survey. The question prompt is as follows:

> The judge is punctual in convening court, keeps business moving, and does an amount of work fair to taxpayers and to other judges.

While many of the questions are double-barreled items, this is triple-barreled. This is problematic, as it forces the respondent to make judgments about which component of the question deserves the most weight. How should an attorney rate a judge who appears to work very hard, but who also was late to court? The question prompt does not provide any guidance, and attorneys will be left to their own sense of which component should weigh most heavily in their evaluation. Abstract and vague language is pervasive in this question as well. The attorneys are left to figure out for themselves what it means to "keep business moving."

The rating scale is also problematic. Attorney respondents are asked to rate judge performance on each dimension as "more than adequate," "adequate," or "not adequate." This scale requires each respondent attorney to concoct an individual assessment of what kind of behavior qualifies as "adequate." In addition, it makes little sense to answer the Question 11 decision prompt with the phrase "not adequate." The rating scale is incongruous with the question prompt.

Elek, Rottman, and Cutler (2012) argue that these problems compromise the validity, reliability, and perceived fairness of the JPE process. Indeed, there are some warning signs in the aggregate respondent data. The inter-dimensional correlation matrix (see Table 2) shows a high alpha score and high bivariate correlations between all of the five dimensions of judicial performance. Three of the dimensions are measured by two or more questions on the survey; these intra-dimensional correlation matrices can be found in Table 3. These correlations and alpha scores are only slightly

[4]The administrative performance scores are measured for only 87 judges, yielding only 311 observations; the Judging the Judges survey does not collect this information for Nevada Supreme Court Justices.

[5]The precise wording and organization of the questions has changed slightly over time. The years column in Table 1 indicates which years the Judging the Judges survey used each of the questions. The Q (Question) column indicates which historical and question formulations make up the scores for each of the questions contained in the analysis. For example, the judge-level scores on Question 1 in the analysis are the average of two questions in the 1998–2000 data, but they are the results of a single question in the 2002–2008 data.

TABLE 1
Questions on the LVRJ Judging the Judges Survey Instrument, 1998–2008

Q	*Years*	*Category*	*Survey Question Text*
1	98–00	Integrity & Impartiality	The judge fairly and impartially weighs all the evidence and arguments of counsel before rendering a decision.
		Integrity & Impartiality	The judge demonstrates familiarity with the pleadings, record, memoranda, and/or briefs.
	02–08	Integrity & Impartiality	The judge demonstrates familiarity with the case record and documents, and fairly weighs all evidence and arguments before rendering a decision.
2	98–00	Legal Ability	The judge's rulings regarding criminal sentencing and contempt are appropriate.
	02–08	Legal Ability	The judge's rulings, whether regarding civil issues, criminal sentencing, or contempt, are appropriate.
3	98–00	Legal Ability	The judge properly applies the rules of procedure and evidence
		Legal Ability	The judge properly applies the law.
	02–08	Legal Ability	The judge properly applies the law, rules of procedure, and rules of evidence.
4	98–08	Communication Skills	The judge clearly explains the basis for his or her decisions.
5	98–08	Integrity & Impartiality	The judge's professional behavior is free from impropriety or the appearance of impropriety.
6	98–08	Integrity & Impartiality	The judge's conduct is free from bias on the basis of race or ethnic origin.
7	98–08	Integrity & Impartiality	The judge's conduct is free from bias on the basis of gender.
8	98–08	Integrity & Impartiality	The judge's conduct is free from bias on the basis of religion.
9	98–08	Integrity & Impartiality	The judge's conduct is free from bias on the basis of parties or attorneys involved in the action.
10	98–08	Administrative Capacity	The judge issues orders, judgments, decrees or opinions without unnecessary delay.
11	98–00	Administrative Capacity	The judge is punctual in convening court and moves proceedings in an appropriately expeditious manner.
		Administrative Capacity	The judge does an amount of work that is fair to taxpayers and to the other judges of the same court.
	02–08	Administrative Capacity	The judge is punctual in convening court, keeps business moving, and does an amount of work fair to taxpayers and to other judges.
12	98–08	Professionalism & Temperament	The judge is courteous.
Ret	98–08		Taking everything into account, would you recommend retaining this judge on the bench?

higher than the inter-dimensional scores. A factor analysis on all twelve questions shows these questions arrayed on a single dimension.[6]

Respondent attorneys generally give very high scores to judges. The response breakdown by question is presented in Figure 1. Table 4 presents descriptive statistics of a combined score for each question, as well as by category. These combined scores were calculated by weighting the aggregate response totals such that a judge with 100 percent "more than adequate" responses

[6]Principal factors method yields a single factor with an Eigenvalue of 10.570, $N = 311$, Likelihood Ratio test = 8137.79, $p = .000$. All twelve questions load onto this factor with factor loadings greater than 0.85.

TABLE 2
Inter-dimensional Correlation Matrix

	Legal Ability	*Integrity & Impartiality*	*Communication Skills*	*Professionalism & Temperament*	*Administrative Capacity*
Legal Ability	1.000				
Integrity & Impartiality	0.938	1.000			
Communication Skills	0.934	0.962	1.000		
Professionalism & Temperament	0.885	0.909	0.892	1.000	
Administrative Capacity	0.860	0.797	0.844	0.684	1.000

Note. Average inter-item covariance = 0.059, $\alpha = 0.968$

would score a 1.0 and a judge with 100 percent "not adequate" responses would score a −1.0. These distributions are presented graphically in Figure 2. The attorney responses are negatively skewed, meaning that they tended to rate the judges quite high across all questions. However, these data do show a significant percentage of judges rated farther than one standard deviation above or below the mean. This indicates that the questions are able to distinguish among judges based on some criteria.

The more important issue is determining which criteria respondents are using to distinguish judges. The sex and race of a judge should have no relationship to that judge's professional performance. But previous research suggests that there are significant gender and race differences on attorney survey measures of judicial performance (Burger 2007; Gill et al. 2011). More appropriate covariates for judicial performance ratings would be objective measures of judicial performance (or their proxies). For this reason, a number of control variables are included in

TABLE 3
Intra-dimensional Correlation Matrices

A. Legal Ability

	Question 2	*Question 3*
Question 2	1.000	
Question 3	0.971	1.000

Average inter-item covariance = 0.077, $\alpha = 0.997$

B. Administrative Capacity

	Question 10	*Question 11*
Question 10	1.000	
Question 11	0.971	1.000

Average inter-item covariance = 0.054, $\alpha = 0.969$

C. Integrity & Impartiality

	Question 1	*Question 5*	*Question 6*	*Question 7*	*Question 8*	*Question 9*
Question 1	1.000					
Question 5	0.906	1.000				
Question 6	0.906	0.919	1.000			
Question 7	0.868	0.893	0.910	1.000		
Question 8	0.882	0.893	0.935	0.916	1.000	
Question 9	0.907	0.957	0.899	0.910	0.901	1.000

Note. Average inter-item covariance = 0.048, $\alpha = 0.980$

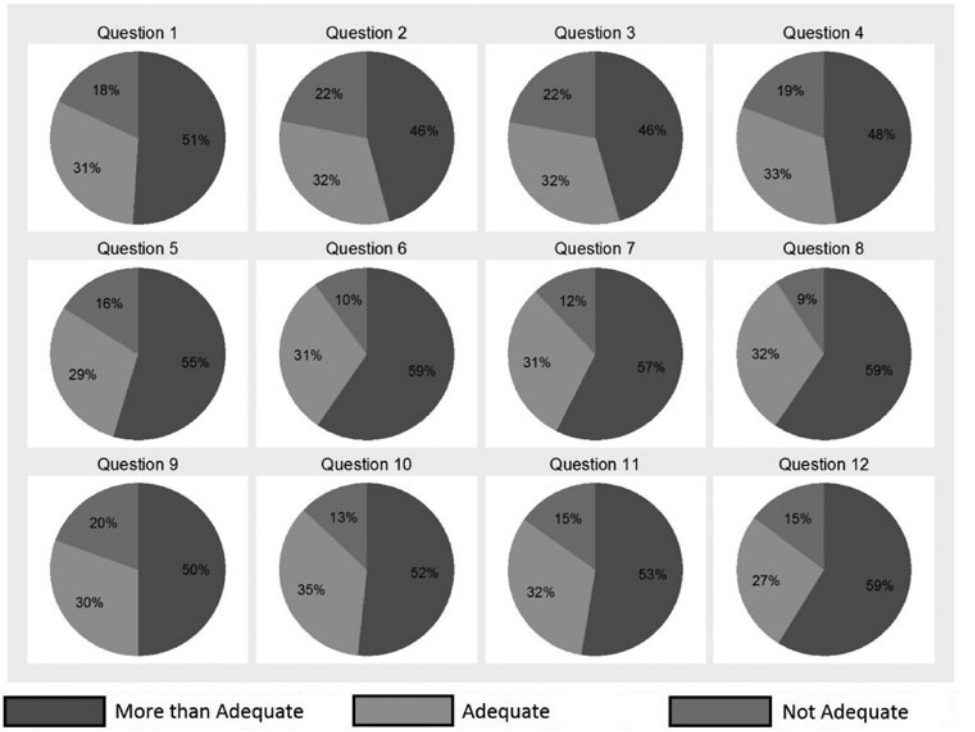

FIGURE 1 Attorney Rating Breakdown by Question.

the multivariate analyses.[7] These variables are detailed in Table 5, and descriptive statistics are presented in Table 6.

A series of pooled OLS regressions isolates the effects of these variables on the attorney ratings. Table 7 presents a series of analyses using various summary indicators as dependent variables. The models all perform quite well, but the results are disconcerting. The column labeled "Retain" shows the effect of the various independent variables on the percentage of attorneys who indicated that the judge should be retained. After controlling for a number of important indicators of judicial performance, women scored nearly 12 points lower out of 100 than men; minority judges scored 21 points lower than white judges. This pattern continues throughout the rest of the analyses. Women and minority judges were significantly less likely to receive "more than adequate" ratings and were significantly more likely to receive "not adequate" ratings. In the various ABA categories, which are scaled here from -1 to $+1$, women and minority judges fared worse than their male and white counterparts across the board.

Table 8 and Table 9 use as the dependent variables the percentage of attorneys rating the judge "more than adequate" and "not adequate," respectively, on each question. This is important

[7]Multicollinearity tests on the set of independent variables show no signs of problematic collinearity. No variance inflation factor is higher than 1.82, and no tolerance is lower than 0.62.

TABLE 4
Combined Score Descriptive Statistics

	N	*Mean*	*Std. Dev.*	*Min*	*Max*	*Number Beyond 1 Standard Deviation*	
						Below	*Above*
Question 1 (II)	350	.33	.27	−.74	.77	33	59
Question 2 (LA)	311	.24	.28	−.79	.74	52	54
Question 3 (LA)	350	.23	.28	−.78	.76	55	63
Question 4 (CS)	350	.29	.27	−.72	.77	47	60
Question 5 (II)	350	.35	.25	−.82	.81	53	75
Question 6 (II)	350	.49	.21	−.53	.85	51	43
Question 7 (II)	350	.45	.22	−.53	.83	56	52
Question 8 (II)	350	.50	.19	−.46	.81	49	43
Question 9 (II)	350	.31	.24	−.72	.75	58	47
Question 10 (AC)	311	.39	.22	−.68	.78	42	50
Question 11 (AC)	311	.37	.25	−.79	.84	42	48
Question 12 (PT)	350	.44	.30	−.70	.92	53	47
Legal Ability	350	.24	.27	−.79	.75	54	61
Integrity & Impartiality	350	.41	.22	−.63	.79	53	52
Communication Skills	350	.29	.27	−.72	.77	53	58
Professionalism & Temperament	350	.44	.30	−.70	.92	53	47
Administrative Capacity	311	.38	.24	−.74	.81	45	45
Retention	350	75.91	14.91	8.00	97.00	56	50

because the Judging the Judges survey reports these percentages to the public. Again, the pattern of sex and race disparity is stark. Women and minorities get significantly fewer "more than adequate" ratings and significantly more "not adequate" ratings compared with their male and white counterparts. Across all of these measures, judge sex and race are more significant and higher magnitude predictors of judicial performance ratings than any of the other independent variables.

That the remaining independent variables perform so poorly in these analyses is another indictment of the criterion validity of the prototypical attorney survey of judicial performance. The effects of these variables are summarized in Table 10. The most direct outside measure of judicial behavior we have is the number of times the judge has been reversed. It seems reasonable to expect that reversals would be strongly related to other measures of performance, especially measures of legal ability. In fact, reversals had little impact on performance ratings. Lower reversal numbers were significant predictors of "more than adequate" ratings on the appropriateness and promptness of the judge's rulings, but high reversal numbers did not increase the percentage of "not adequate" ratings.

The proxy measures of judicial ability were also relatively weak predictors of attorney ratings. Although previous research suggests that more experience on the bench should increase judicial performance (Epstein et al. 2003; Haire 2001), the attorney ratings declined markedly as judicial experience increased. The prestige of the judge's legal education was generally associated with higher attorney ratings. In the summary measures, higher-prestige law schools were associated with higher ratings in the categories of integrity and impartiality and administrative capacity, but

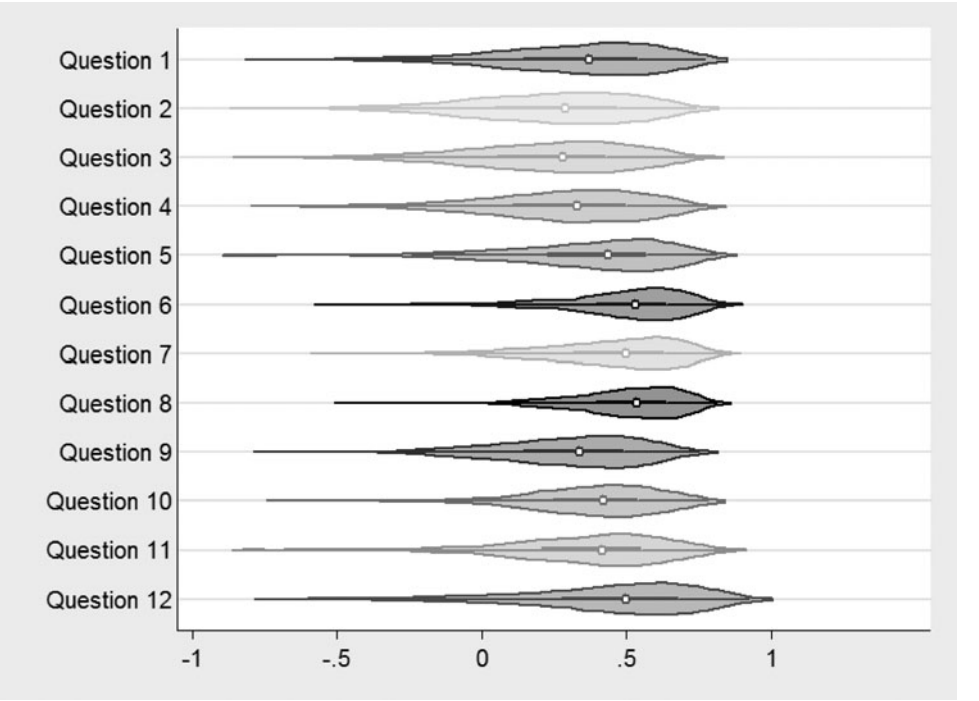

FIGURE 2 Distribution of Combined Scores by Question.

not measures of legal ability or communication skills. Judges who were initially appointed to the bench through Nevada's merit plan for interim appointments were indistinguishable from their elected counterparts in the attorney ratings. State supreme court justices performed similarly to judges on other courts.

The models also contain a variety of measures that speak to issues of professionalism, temperament, integrity, and impartiality. Judges who faced disciplinary action had lower scores on the administrative capacity questions, but there was no overall effect of disciplinary outcomes on the integrity and impartiality score. Worse disciplinary outcomes were associated with a slight increase in the percentage of "not adequate" ratings on race or ethnic bias and a decrease in the "more than adequate" ratings on the administrative capacity questions.

Judges who were involved in official election disputes fared worse than other judges in almost all of the summary categories. These disputes made the judges less likely to get "more than adequate" ratings from attorneys, although they did not result in more "not adequate" ratings across the board. The number of scandals reported in the media had no effect on the summary indicators. However, judges with more scandals fared worse than other judges on two questions dealing with impartiality and integrity. Finally, being identified publicly as a partisan had little effect on the overall scores, although it seemed to provide some protection against "not adequate" ratings in a couple of the performance measures.

TABLE 5
Description of Independent Variables

Female	1 if judge is female, 0 if judge is male
Minority	1 if judge is a member of a racial or ethnic minority group, 0 if judge is white (Note: There were too few minority judges to break this category into different types of minority groups. However, the results were robust to alternative specifications excluding various sub-groups.)
Reversals	The number of times the judge has been reversed by a higher court during the evaluation period divided by the number of cases appealed. (Note: Alternative specifications, including raw reversals and court-standardized reversal rates yielded similar results.)
Experience	number of years the judge has spent in any judicial position at the time of the evaluation (Note: Alternative specifications, including nonlinear specifications, yielded similar results.)
Law School	The 2012 *US News & World Report* rankings for the judge's law school alma mater (1 = Top 14; 2 = 15th–50th; 3 = 51st–100; 4 = 101st-ranked schools; 5 = unranked school; 6 = did not attend)
Appointed	1 if the judge was originally appointed via Nevada's interim merit selection plan, 0 if the judge was initially elected
Supreme Ct.	1 if the judge was serving on the Nevada Supreme Court during the evaluation period, 0 if the judge was serving on some other court.
Discipline	The most severe outcome of disciplinary action from the Commission on Judicial Discipline against the judge during the review period. (0 = no complaint filed; 1 = complaint dismissed; 2 = required course; 3 = required course and public apology; 4 = public reprimand; 5 = public reprimand and fine; 6 = censure, required course, and fine; 7 = removal from bench; 8 = removal from bench and permanently barred from holding public office)
Election	1 if the judge was involved in an election complaint in front of the Standing Committee on Judicial Ethics and Election Practices, 0 if the judge was not involved in such a complaint. (Note: Alternative specifications indicating whether the judge was the defendant or the accused in an election dispute yielded similar results.)
Scandals	The number of *Las Vegas Review-Journal* articles linking the judge to a scandal during the review period. (Note: Alternative specifications, including weighting the number of scandal-related articles by the number of neutral or neutral and positive stories yielded similar results.)
Partisan	1 if judge has been associated publicly with a political party, 0 if not (Note: Alternative specifications, including whether the judge has been identified with the Republican or Democratic parties, yielded similar results.)

TABLE 6
Descriptive Statistics for Independent Variables

	N	*Mean*	*Std. Dev.*	*Min*	*Max*
Female	94	.34	.48	0	1
Minority	94	.06	.25	1	0
Reversals	94	3.92	4.56	0	20.79
Experience	94	7.87	6.70	1	27
Law School	94	2.15	1.23	0	5
Appointed	94	.40	.49	0	1
Sup. Court	94	.11	.29	0	1
Discipline	94	.54	1.72	0	8
Election	94	.09	.28	0	1
Scandals	94	.21	.41	0	1
Partisan	94	.59	.50	0	1

Note. For ease of interpretation, these descriptive statistics are collapsed by judge.

TABLE 7
Summary Indicators (Pooled OLS with Huber-White Sandwich Standard Errors)

	Retain	*MA*	*NA*	*All*	*LA*	*II*	*CS*	*PT*	*AC*
Female	−11.606***	−9.687***	7.835***	−0.175***	−0.234***	−0.144**	−0.187**	−0.148**	−0.249***
	(3.123)	(2.724)	(2.071)	(0.047)	(0.059)	(0.046)	(0.059)	(0.057)	(0.049)
Minority	−20.803***	−16.556**	15.485**	−0.320**	−0.347***	−0.295**	−0.395**	−0.414**	−0.338***
	(5.927)	(5.424)	(5.061)	(0.103)	(0.103)	(0.098)	(0.134)	(0.165)	(0.084)
Reversals	−0.195	−0.387	0.051	−0.004	−0.008	−0.003	−0.006	−0.005	−0.008
	(0.332)	(0.304)	(0.214)	(0.005)	(0.006)	(0.005)	(0.018)	(0.006)	(0.006)
Experience	−0.327*	−0.417**	0.219*	−0.006**	−0.006*	−0.006*	−0.009**	−0.009*	−0.007**
	(0.158)	(0.150)	(0.107)	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)	(0.002)
Law School	−2.005*	−2.290**	1.338*	−0.036*	−0.034	−0.041**	−0.032	−0.060**	−0.012
	(0.993)	(0.86)	(0.680)	(0.005)	(0.018)	(0.015)	(0.018)	(0.020)	(0.043)
Appointed	2.524	2.180	−1.709	0.039	0.049	0.045	0.042	0.002	0.045
	(2.711)	(2.441)	(1.752)	(0.041)	(0.050)	(0.040)	(0.049)	(0.052)	(0.04.)
Supreme Ct.	−2.582	2.914	−0.196	0.032	0.008	0.006	0.041	0.068	—
	(3.466)	(3.069)	(2.104)	(0.051)	(0.065)	(0.050)	(0.060)	(0.061)	
Discipline	−1.612	−1.070	1.286	−0.024	−0.019	−0.023	−0.014	−0.020	−0.031**
	(0.947)	(0.765)	(0.685)	(0.065)	(0.016)	(0.015)	(0.016)	(0.021)	(0.012)
Election	−8.848*	−9.268**	5.532	−0.148*	−0.143*	−0.124*	−0.188*	−0.354*	−0.092
	(4.472)	(3.601)	(3.066)	(0.065)	(0.070)	(0.061)	(0.074)	(0.148)	(0.061)
Scandals	−3.111	−4.568	1.848	−0.064	−0.076	−0.065	−0.087	−0.061	−0.045
	(3.291)	(2.886)	(2.106)	(0.049)	(0.062)	(0.048)	(0.031)	(0.043)	(0.054)
Partisan	3.294	1.010	−2.693*	0.037	0.036	0.039	0.026	0.034	0.062
	(2.574)	(2.355)	(0.107)	(0.040)	(0.048)	(0.387)	(0.016)	−0.050	(0.043)
Constant	88.801***	68.005***	7.872***	0.601***	0.494***	0.628***	0.549***	0.771***	0.582***
	(3.347)	(3.191)	(2.223)	(0.053)	(0.068)	(0.049)	(0.065)	(0.058)	(0.055)
N (judges)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	311 (86)
F Test	10.23***	9.92***	8.77***	9.93***	9.54***	9.75***	8.90***	8.44***	10.52***
R^2	.33	.37	.35	.37	.35	.35	.35	.36	.42
Adjusted R^2	.31	.35	.33	.35	.33	.33	.32	.34	.40
Root MSE	12.359	11.063	8.254	0.186	0.225	0.180	0.220	0.243	0.182

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

TABLE 8
More than Adequate Ratings (Pooled OLS with Huber-White Sandwich Standard Errors)

	Q1 (II)	*Q2 (LA)*	*Q3 (LA)*	*Q4 (CS)*	*Q5 (II)*	*Q6 (II)*	*Q7 (II)*	*Q8 (II)*	*Q9 (II)*	*Q10 (AC)*	*Q11 (AC)*	*Q12 (PT)*
Female	−10.594**	−12.552***	−10.969***	−9.408**	−8.231**	−8.284**	−8.926***	−7.534**	−7.505**	−13.149***	−14.735***	−9.056**
	(3.541)	(3.281)	(3.192)	(3.286)	(3.062)	(2.538)	(2.543)	(2.521)	(2.615)	(2.555)	(2.862)	(3.442)
Minority	−19.223**	−17.204***	−17.452**	−19.452**	−14.710*	−19.664***	−13.954**	−13.525**	−12.664**	−18.011***	−18.618***	−21.866**
	(6.395)	(5.237)	(5.774)	(6.233)	(6.276)	(5.530)	(4.867)	(4.597)	(4.635)	(4.367)	(4.763)	(8.642)
Reversals	−0.284	−0.850*	−0.606	−0.512	−0.330	−0.084	−0.367	−0.119	−0.542	−0.870*	−0.562	−0.458
	(0.405)	(0.387)	(0.393)	(0.356)	(0.349)	(0.256)	(0.275)	(0.248)	(0.301)	(0.632)	(0.381)	(0.344)
Experience	−0.543***	−0.367*	−0.428**	−0.537***	−0.465**	−0.333*	−0.277	−0.283*	−0.381*	−0.435**	−0.537***	−0.561*
	(0.163)	(0.165)	(0.154)	(0.156)	(0.179)	(0.145)	(0.144)	(0.141)	(0.161)	(0.141)	(0.155)	(0.251)
Law School	−1.986	−2.047*	−2.155*	−2.148*	−2.755**	−2.299**	−2.932***	−2.377**	−2.644**	−1.144	−1.220	−3.693***
	(1.066)	(1.001)	(1.010)	(1.022)	(0.913)	(0.792)	(0.875)	(0.792)	(0.817)	(0.784)	(0.903)	(1.129)
Appointed	1.797	2.682	2.820	2.499	2.667	2.782	3.902	2.118	2.024	1.941	1.996	0.169
	(2.993)	(3.052)	(2.821)	(2.810)	(2.657)	(2.260)	(2.365)	(2.267)	(2.425)	(2.424)	(2.646)	(3.129)
Supreme Ct.	3.068	—	2.876	2.938	−0.294	1.178	1.352	0.460	0.891	—	—	3.621
	(4.053)		(3.804)	(3.547)	(3.970)	(2.611)	(2.652)	(2.527)	(3.318)			(3.674)
Discipline	−0.939	−0508	−0.599	−0.490	−1.579	−1.108	−1.342	−0.857	−0.995	−1.184*	−1.657*	−0.969
	(.0961)	(0.781)	(0.870)	(0.869)	(0.945)	(0.0784)	(0.843)	(0.769)	(0.799)	(0.558)	(0.683)	(1.218)
Election	−8.839*	−9.132*	−7.871*	−10.660**	−8.624*	−7.923*	−8.100	−7.237*	−7.993*	−6.934*	−7.794*	−19.639**
	(3.720)	(4.671)	(3.654)	(3.948)	(3.538)	(3.471)	(4.393)	(3.641)	(3.357)	(2.800)	(3.836)	(7.455)
Scandals	−5.256	−5.684	−5.285	−5.451	−6.446*	−3.294	−3.079	−3.177	−5.767*	−3.721	−3.745	−4.656
	(3.788)	(3.830)	(3.568)	(3.501)	(3.109)	(2.673)	(2.697)	(2.773)	(2.768)	(2.932)	(3.445)	(4.501)
Partisan	0.746 2.932	2.003	0.074	0.079	1.263	2.082	1.337	2.040	0.872	1.482	2.988	1.409
		(2.756)	(2.749)	(2.734)	(2.561)	(2.138)	(2.274)	(2.169)	(2.307)	(2.291)	(2.575)	(3.021)
Constant	66.575***	61.998***	61.735***	64.462***	70.741***	71.346***	71.057***	70.603***	65.673***	66.955***	67.872***	79.747***
	(3.968)	(3.700)	(3.922)	(3.720)	(3.327)	(2.832)	(2.922)	(2.830)	(3.096)	(3.196)	(3.447)	(3.552)
N (judges)	350 (94)	311 (87)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	311 (87)	311 (87)	350 (94)
F Test	8.05***	8.71***	7.83***	8.51***	11.23***	8.90***	9.30***	6.47***	11.16***	13.26***	11.49***	9.29***
R^2	.32	.36	.33	.34	.35	.34	.35	.28	.35	.42	.42	.36
Adjusted R^2	.29	.34	.31	.32	.33	.32	.33	.26	.33	.40	.40	.34
Root MSE	13.197	12.398	12.658	12.438	12.208	10.759	11.052	10.695	11.031	10.764	11.392	14.214

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

TABLE 9
Not Adequate Ratings (Pooled OLS with Huber-White Sandwich Standard Errors)

	Q1 (II)	*Q2 (LA)*	*Q3 (LA)*	*Q4 (CS)*	*Q5 (II)*	*Q6 (II)*	*Q7 (II)*	*Q8 (II)*	*Q9 (II)*	*Q10 (AC)*	*Q11 (AC)*	*Q12 (PT)*
Female	10.029***	12.832***	12.560***	9.288***	5.679**	4.628**	5.866***	3.332*	5.535**	9.755***	12.063***	5.784*
	(2.796)	(3.168)	(3.031)	(2.737)	(2.241)	(1.547)	(1.785)	(1.536)	(2.157)	(2.236)	(2.591)	(2.328)
Minority	18.420**	19.170***	19.612***	20.027**	14.151**	20.702***	10.015*	7.755*	12.044**	13.042***	17.894***	19.519*
	(5.925)	(4.788)	(5.494)	(7.312)	(5.184)	(6.090)	(4.065)	(3.346)	(4.823)	(3.967)	(4.498)	(7.945)
Reversals	0.049	0.274	0.175	0.107	−0.080	−0.176	0.230	−0.777	0.245	0.273	−0.133	0.205
	(0.287)	(0.379)	(0.304)	(0.278)	(0.238)	(0.175)	(0.215)	(0.165)	(0.240)	(0.301)	(0.304)	(0.292)
Experience	0.289*	0.243	0.208	0.332*	0.295*	0.144	0.086	0.110	0.254*	0.128	0.306*	0.357*
	(0.127)	(0.138)	(0.125)	(0.136)	(0.135)	(0.086)	(0.095)	(0.099)	(0.129)	(0.094)	(0.126)	(0.179)
Law School	1.059	1.152	1.398	1.030	2.019**	1.368*	2.022**	1.244*	2.189**	−0.029	−0.033	2.294**
	(0.878)	(0.902)	(0.926)	(0.861)	(0.787)	(0.555)	(0.691)	(0.564)	(0.767)	(0.606)	(0.744)	(0.872)
Appointed	−1.910	−2.137	−2.600	−1.734	−2.051	−1.320	−3.683*	−0.882	−1.173	−2.698	−2.315	−0.034
	(2.228)	(2.616)	(2.423)	(2.221)	(1.944)	(1.275)	(1.656)	(1.433)	(1.989)	(1.827)	(2.193)	(2.180)
Supreme Ct.	−0.949	—	1.427	−1.176	2.502	0.522	−0.602	0.919	0.804	—	—	−3.174
	(2.710)		(2.852)	(2.590)	(3.048)	(4.460)	(1.693)	(1.645)	(2.644)			(2.494)
Discipline	1.167	1.341	1.020	0.844	1.407	1.180*	1.541	0.738	0.804	1.349*	2.042**	1.070
	(0.864)	(0.764)	(0.885)	(0.837)	(0.783)	(0.587)	(0.791)	(0.621)	(2.643)	(0.587)	(0.670)	(0.996)
Election	6.254*	6.781	6.004	8.175*	4.780	2.961	2.966	2.805	6.118*	1.682	2.085	15.773*
	(3.074)	(4.488)	(3.140)	(3.688)	(2.896)	(2.251)	(3.251)	(1.991)	(2.953)	(2.916)	(3.492)	(7.472)
Scandals	2.753	3.013	2.4778	3.201	3.631	0.164	0.583	1.179	3.700*	0.503	1.079	1.410
	(2.753)	(3.398)	(2.945)	(2.699)	(2.256)	(1.798)	(2.065)	(2.091)	(2.953)	(2.234)	(2.674)	(3.211)
Partisan	−3.015	−4.922*	−3.355	−2.550	−2.417	−2.203	−2.412	−2.839*	−2.117	−3.082	−4.922*	−1.954
	(2.263)	(2.491)	(2.397)	(2.264)	(1.948)	(1.242)	(1.604)	(1.385)	(44.928)	(1.910)	(2.217)	(2.089)
Constant	9.321***	12.200***	12.262***	9.550***	6.397**	4.432**	5.240**	5.092**	8.613***	8.591***	9.899***	2.627
	(2.846)	(3.103)	(3.181)	(2.895)	(2.373)	(1.667)	(1.917)	(1.783)	(2.381)	(2.246)	(2.713)	(2.451)
N (judges)	350 (94)	311 (87)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	350 (94)	311 (87)	311 (87)	350 (94)
F Test	8.21***	10.25***	10.22***	8.41***	8.35***	6.47***	7.85***	3.39***	9.87***	5.93***	8.25***	6.92***
R^2	.32	.35	.33	.32	.32	.43	.35	.21	.30	.33	.40	.32
Adjusted R^2	.30	.32	.31	.30	.29	.41	.33	.19	.28	.30	.38	.30
Root MSE	10.404	11.365	11.343	10.531	9.442	6.671	7.794	6.656	9.504	8.155	9.211	10.919

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

TABLE 10
Summary of Independent Variable Effects

Female	Female judges scored significantly lower than male judges across all of the possible measures. Women judges fared particularly poorly in measures of Administrative Capacity and Legal Knowledge.
Minority	Minority judges scored significantly lower than white judges across all of the possible measures.
Reversals	Reversal rates had almost no impact on performance ratings. Lower reversal rates were significant predictors of "more than adequate" ratings in one Legal Ability question and one Administrative Capacity question.
Experience	Each additional year of experience led to a significantly lower score across nearly every measure.
Law School	Higher prestige law schools were associated with generally higher scores. Law school prestige increased scores on summary measures of Integrity and Impartiality as well as Administrative Capacity, but not on measures of Legal Ability or Communication Skills.
Appointed	Judges initially selected through Nevada's merit system for interim appointments were indistinguishable from those initially elected to the bench.
Supreme Ct.	State supreme court justices were indistinguishable from judges in other courts.
Discipline	The most serious outcomes on disciplinary action were associated with lower scores on Administrative Capacity questions, but there was no overall effect on the Integrity and Impartiality score. Worse disciplinary outcomes were also associated with a slight increase in the number of "not adequate" ratings on Question 6, which deals with Impartiality and Integrity.
Election	Judges who were involved in election disputes fared worse than other judges in almost all of the summary categories. These disputes made the judges less likely to get "more than adequate" ratings from attorneys, although they did not result in more "not adequate" ratings across the board.
Scandals	The number of scandals reported in the media had no effect on the summary indicators. Judges with more scandals fared worse than other judges on Questions 5 and 9, both of which deal with Impartiality and Integrity.
Partisan	Being identified in the media as a partisan had little effect, but seemed to provide some protection against "not adequate" ratings in a couple of the performance measures.

DISCUSSION

Unfortunately, the results presented here suggest that there is significant cause for concern about JPE attorney surveys. The sex and race disparities in the Judging the Judges survey act as a thumb on the scales, systematically disadvantaging groups that have been traditionally underrepresented on the bench. There is not a single category of questions that escapes this problem; the effects of judge sex and race are significant, large, and consistent across all of the dimensions of judicial performance evaluated by the Judging the Judges survey.

There is reason to believe that the source of this systematic gap in scores is implicit bias. According to implicit social cognition theory, JPE attorney surveys are prime candidates for implicit race and gender bias. Indeed, this kind of bias flourishes in performance appraisal instruments for gender and race stereotyped jobs. Such bias is also fed by the kind of validity and reliability problems that are known to plague attorney surveys. While the lack of individual respondent-level data makes a full reliability and validity analysis impossible, the patterns in the scores derived from this JPE instrument are consistent with what we would expect to see from an instrument with significant reliability and validity problems.

Even setting aside the large problem of sex and race bias, the other measures of judicial performance often do not have the theoretically expected effect on measures of judicial performance. The inter- and intra-dimensional correlation matrices suggest that the various measures do not

seem to isolate the dimensions of judicial performance. That reversals are largely insignificant in the multivariate models is worrisome. That judicial disciplinary outcomes are not strongly related to measures of integrity and impartiality is problematic. That scores go down consistently with increased experience on the bench is troubling.

The bad news does not end there. Despite good intentions, it is not clear that JPEs are designed and administered in a way that guarantees fair and reliable results. Most jurisdictions that conduct JPEs follow the American Bar Association's "Guidelines for the Evaluation of Judicial Performance" (American Bar Association 1985). A recent update to the ABA's Guidelines (American Bar Association 2005) acknowledges the potential for bias in judicial performance evaluations. But the paper dismisses the problem out of hand on the basis of the fact that the Guidelines recommend behavior-based evaluation questions:

> An additional benefit of behavior-based evaluation instruments is that questionnaire items reduce subjectivity in assessments of judicial performance, thus limiting the potential for gender and other biases to influence responses. In a behavior-based evaluation instrument, items relate to judges' actual behaviors rather than characterizations of judges' actions as proper or improper. (American Bar Association 2005, 14)

The performance appraisal research generally supports such a claim. But the problem lies in the fact that the surveys as implemented in the states do not systematically reach this ideal.

The questions on the Judging the Judges survey are derived from the ABA Guidelines, and the questions in the survey are very similar to those questions used in model attorney surveys (Elek et al. 2012) and existing state-sponsored JPE programs across the country (Gill et al. 2011). Take, for example, the Colorado JPE survey. Colorado's system of judicial performance evaluation is one of the most studied and well-funded in the nation (Brody 2008b). The Executive Director of the Institute for the Advancement of the American Legal System (IAALS) says of Colorado's survey:

> All those individuals complete survey questionnaires that focus on things such as the following: was the judge prepared when he or she showed up on the bench; was the judge respectful of the people in the courtroom; did the judge move the docket along efficiently; was the judge timely in his or her rulings and were those rulings clear and understandable? In other words, the questions focus on process, not on outcome. (Kourlis 2010, 767)

But these are exactly the same kinds of questions that we find on the Judging the Judges survey. And many of the other characteristics of the Judging the Judges survey are also common to JPEs in other states, including the primacy of the survey in the in the JPE process (Pelander 1998), the survey design problems (Elek et al. 2012), and the low response rates (Brody 2000). As such, there is good reason to believe that the very real problems identified here are being replicated across the country.

None of this would be a problem except for the fact that the results of these judicial performance evaluations may actually matter. As Judge Kourlis explains, surveys of voters found that "if voters knew that we had a judicial performance evaluation system, they would use the information and they would trust it" (Kourlis 2010, 768). She says of these voters:

> They trusted the fact that the judicial performance commissioners were looking at the right data and making good decisions; and, thus, the voters could turn to the data and those decisions to guide them. That is what an informed vote should look like. (Kourlis 2010, 768)

The results of this analysis do not show conclusively that the problems plaguing the Judging the Judges survey are present to the same degree in all JPE surveys. There are some unique characteristics of the Clark County survey that may serve to exacerbate the findings of bias. For example, the fact that it is not an official, state-sponsored survey may cause respondent attorneys to take the process less seriously. There may also be some characteristics of the legal culture in Nevada influencing the results. However, the similarities between the Judging the Judges survey and other JPE surveys are striking. Additional research is necessary to rule out the possibility that the deficiencies in design and execution demonstrated by Elek et al. (2012) lead to the kind of results identified in this analysis.

In short, following the ABA's Guidelines is probably not enough to prevent serious bias from poisoning judicial performance evaluation scores. This is a particularly poor outcome, as it means subjecting many judges to state-sponsored evaluations that are systematically biased against women and minorities. This analysis raises real questions about the overall validity and reliability of the JPE survey as a measurement of judicial performance. Perhaps it is, as some Colorado judges suspect, simply "a popularity contest" (IAALS 2006a). While some lament the fact that many voters are unaware of the judicial performance evaluation data when they make their decisions (Kourlis 2010; White 2009), perhaps instead we should be relieved. All of this does not mean that the entire enterprise of evaluating judicial performance should be abandoned; however, we must do better than this.

REFERENCES

American Bar Association, Special Commission on Evaluation of Judicial Performance. 1985. *Guidelines for the Evaluation of Judicial Performance*. Washington, DC: American Bar Association.

———. 2005. *Guidelines for the Evaluation of Judicial Performance with Commentary*. Chicago: American Bar Association. Available at http://www.americanbar.org/content/dam/aba/publications/judicial_division/aba_blackletterguidelines_jpe_wcom.authcheckdam.pdf

Aynes, Richard L. 1981. "Evaluation of Judicial Performance: A Tool for Self-Improvement." *Pepperdine Law Review* 8(2):255–312.

Banks, Cristina G., and Kevin R. Murphy. 1985. "Toward Narrowing the Research-Practice Gap in Performance Appraisal." *Personnel Psychology* 38(2):335–45.

Bartlett, F. C. 1932. *Remembering*. New York: MacMillan.

Bautista, Rene. 2010. "An Overlooked Approach in Survey Research: Total Survey Error." In *Handbook of Survey Methodology for the Social Sciences*, edited by L. Gideon, 37–49. New York: Springer.

Bernardin, H. John, Robert Konopaske, and Christine M. Hagan. 2012. "A Comparison of Adverse Impact Levels Based on Top-Down, Multisource, and Assessment Center Data: Promoting Diversity and Reducing Legal Challenges." *Human Resource Management* 51(3):313–341.

Bernick, E. Lee, and David J. Pratto. 1995. "A Behavior-based Evaluation Instrument for Judges." *Justice System Journal* 18(2):173–184.

Boatright, Robert G., and Kevin M. Esterling. 2000. "Methodological Issues in the Study of Judicial Election: A Critique of Empirical Research and Suggestions for Future Directions." In *Research on Judicial Selection 1999*, 73–110. Chicago, IL: American Judicature Society.

Borman, Walter C. 1978. "Exploring Upper Limits of Reliability and Validity in Job Performance Ratings." *Journal of Applied Psychology* 63 (2):135–44.

Bowman, Cynthia Grant. 1998. "Bibliographical Essay: Women and the Legal Profession." *American University Journal of Gender, Social Policy and Law* 7:149.

Brody, David C. 2000. "Judicial Performance Evaluations by State Governments: Informing the Public While Avoiding the Pitfalls." *Justice System Journal* 21:333–56.

———. 2003. "The Relationship Between Judicial Performance Evaluations and Judicial Elections." *Judicature* 87:168–92.
———. 2008a. *Pierce County Superior Court Judicial Performance Evaluation: Final Report*. Spokane: Washington State University Spokane.
———. 2008b. "The Use of Judicial Performance Evluation to Enhance Judicial Accountability, Judicial Independence, and Public Trust." *Denver University Law Review* 86(1):1–42.
———. 2012. *2012 Judicial Evaluation Survey: Evaluations of the Judges of the King County Superior Court*. Seattle, WA: King County Bar Association.
Brody, David C., and Nicholas Lovrich. 2009. "Hearing and Mediation Judges of the Washington Board of Industrial Insurance Appeals: Judicial Performance Evaluation Final Report." Pullman: Washington State University.
Burger, Gary K. 2007. *Attorney's Ratings of Judges: 1998–2006*. Mound City, MO: Report to the Mound City Bar.
Cann, Damon M. 2006. "Beyond Accountability and Independence: Judicial Selection and State Court Performance." *Judicature* 90:226.
Carnes, Molly, Stacie Geller, Eve Fine, Jennifer Sheridan, and Jo Handelsman. 2005. "NIH Director's Pioneer Awards: Could the Selection Process Be Biased against Women?" *Journal of Women's Health* 14(8):684–92.
Chauvin, L. Stanley. 1989. "Judicial-Performance Evaluation." *American Bar Association Journal* 75:8.
Choi, Stephen J., Mitu Gulati, and Eric A. Posner. 2009. "Judicial Evaluations and Information Forcing: Ranking State High Courts and their Judges." *Duke Law Journal* 58:1313.
Cleeremans, A. 2003. "Implicit Learning Models." In *Encyclopedia of Cognitive Science*, edited by L. Nadel, 491–9. New York: Nature Publishing Group.
Clydesdale, Timothy T. 2004. "A Forked River Runs Through Law School: Toward Understanding Race, Gender, Age, and Related Gaps in Law School Performance and Bar Passage." *Law & Social Inquiry* 29(4):711–69.
Coontz, Phyllis D. 1995. "Gender Bias in the Legal Profession: Women "See" It, Men Don't." *Women & Politics* 15(2):1–22.
Daley, Dennis M. 1992. *Performance Appraisal in the Public Sector: Techniques and Applications*. Westport, CT: Quorum Books.
Dubois, Philip L. 1980. *From Ballot to Bench: Judicial Elections and the Quest for Accountability*. Austin: University of Texas Press.
Durham, Christine M. 2000. "Gender and Professional Identity: Unexplored Issues in Judicial Performance Evaluation." *Judges' Journal* 39(2):13–6.
Elek, Jennifer K., David B. Rottman, and Brian L. Cutler. 2012. "Judicial Performance Evaluation: Steps to Improve Survey Process and Measurement." *Judicature* 95:65–75.
Epstein, Lee, Jack Knight, and Andrew D. Martin. 2003. "The Norm of Prior Judicial Experience and Its Consequences for Career Diversity on the U.S. Supreme Court." *California Law Review* 91(4):903–65.
Esterling, Kevin M. 1999. "Judicial Accountability the Right Way." *Judicature* 82:206–15.
Esterling, Kevin M., and Kathleen M. Sampson. 1998. *Judicial Retention Evaluation Programs in Four States: A Report with Recommendations*. Chicago, IL: American Judicature Society.
Feeney, Floyd. 1987. "Evaluating Trial Court Performance." *Justice System Journal* 12 (1):148–70.
Fiske, Susan T., Donald N. Bersoff, Eygene Borgida, Kay Deaux, and Madeline E. Heilman. 1991. "Social Science Research on Trial: Use of Sex Stereotyping Research in *Price Waterhouse v. Hopkins*." *American Psychologist* 46(10):1049–1160.
Gill, Rebecca D., Sylvia R. Lazos, and Mallory M. Waters. 2011. "Are Judicial Performance Evaluations Fair to Women and Minorities? A Cautionary Tale from Clark County, Nevada." *Law & Society Review* 45(3):731–59.
Gordon, J. Cunyon. 2003. "Painting by Numbers: "And, Um, Let's Have a Black Lawyer Sit at Our Table." *Fordham Law Review* 71(4):1257.
Griffin, Jacqueline R. 1994. "Judging the Judges." *Litigation* 21(3):5.
Haire, Susan B. 2001. "Rating the Ratings of the American Bar Association Standing Committee on Federal Judiciary." *Justice System Journal* 22(1):1–18.
Hall, William K. 1985. *Judicial Retention Elections: Do Bar Association Polls Increase Voter Awareness?* Urbana: Institute of Government and Public Affairs, University of Illinois.
Hall, William K., and Larry T. Aspin. 1987. "The Roll-Off Effect in Judicial Retention Elections." *Social Science Journal* 24(4):415–27.

Heilman, Madeline E. 1983. "Sex Bias in Work Settings: The Lack of Fit Model." *Research in Organizational Behavior* 5:269–98.

Hekman, David R., Karl Aquino, Bradley P. Owens, Terence R. Mitchell, Pauline Schilpzand, and Keith Leavitt. 2010. "An Examination of Whether and How Racial and Gender Biases Influence Customer Satisfaction." *Academy of Management Review* 53(2):238–64.

Hojnacki, Marie, and Lawrence Baum. 1992. "'New-Style' Judicial Campaigns and the Voters: Economic Issues and Union Members in Ohio." *Western Political Quarterly* 45:921–48.

Hopkins, A. D. 2012. "Judging the Judges: How Lawyers Rate Judges." *Las Vegas Review-Journal*. Available at http://www.reviewjournal.com/news/crime-courts/judging-judges-how-lawyers-rate-judges

IAALS. 2006a. "Shared Expectations: Judicial Accountability in Context." Denver, CO: Institute for the Advancement of the American Legal System.

———. 2006b. "Transparent Courthouse: A Blueprint for Judicial Performance Evaluation." Denver, CO: Institute for the Advancement of the American Legal System.

Kang, Jerry, and Mahzarin Banaji. 2006. "Fair Measures: A Behavioral Realist Revision of 'Affirmative Action'." *California Law Review* 94:1063.

Knowlton, Natalie, and Malia Reddick. 2012. *Leveling the Playing Field: Gender, Ethnicity, and Judicial Performance Evaluation*. Denver, CO: Institute for the Advancement of the American Legal System. Available at http://iaals.du.edu/images/wygwam/documents/publications/IAALS_Level_the_Playing_Field_FINAL.pdf

Kourlis, Rebecca. 2010. "Judicial Performance Evaluation." *Wayne Law Review* 56:765.

Kourlis, Rebecca Love, and Jordan M. Singer. 2007. "Using Judicial Performance Evaluations to Promot Judicial Accountability." *Judicature* 90 (5):200–07.

Krieger, Linda Hamilton. 1995. "The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity." *Stanford Law Review* 47(6):1161–248.

Landy, Frank J., and James Farr. 1980. "Performance Rating." *Psychological Bulletin* 87(2):72–107.

Lee, Audrey J. 2005. "Unconscious Bias Theory in Employment Discrimination Litigation." *Harvard Civil Rights-Civil Liberties Law Review* 40 (481–503):481.

Martin, David C., and Kathryn M. Bartol. 1988. "Training the Raters: A Key to Effective Preformance Appraisal." In *Performance Evaluation: An Essential Management Tool*, edited by C. S. Becker, 192–201. Washington, DC: International City Management Association.

Morrisey, George L. 1983. *Performance Appraisals in the Public Sector: Key to Effective Supervision*. Reading, MA: Addison-Wesley.

Moynihan, Donald P., Sergio Fernandez, Soonhee Kim, Kelly M. LeRoux, Suzanne J. Piotrowski, Bradley E. Wright, and Kaifeng Yang. 2011. "Performance Regimes Amidst Governance Complexity." *Journal of Public Administration Research and Theory* 21(1):141–55.

Olson, Susan M. 2001. "Voter Mobilization in Judicial Retention Elections: Performance Evaluations and Organized Opposition." *Justice System Journal* 22:263–86.

Olson, Susan M., and Christina Batjer. 1999. "Competing Narratives in a Judicial Retention Election: Feminism versus Judicial Independence." *Law & Society Review* 33(1):123–60.

Pelander, A. John. 1998. "Judicial Performance Review in Arizona: Goals, Practical Effects and Concerns." *Arizona State Law Journal* 30:643–758.

Prowse, Peter, and Julie Prowse. 2009. "The Dilemma of Performance Appraisal." *Measuring business excellence* 13(4):69–77.

Puiszis, Steven M. 2011. *Without Fear of Favor in 2011: A New Decade of Challenges to Judicial Independence and Accountability*. Chicago, IL: Defense Research Institute.

Rhode, Deborah L. 2001. *The Unfinished Agenda: Women and the Legal Profession*. Chicago, IL: ABA Commission on Women in the Profession.

Shafritz, Jay M., Albert C. Hyde, and David H. Rosenbloom. 1986. *Personnel management in government: Politics and process,* 3rd ed. New York: M. Dekker.

Singer, Jordan M. 2007. "Knowing is Half the Battle: A Proposal for Prospective Performance Evaluations in Judicial Elections." *University of Arkansas Little Rock Law Review* 29:725.

Smith, Patricia Cain, and Lorne M Kendall. 1963. "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales." *Journal of Applied Psychology* 47(2):149.

Sterling, Joyce S. 1993. "The Impact of Gender Bias on Judging: Survey of Attitudes Toward Women Judges." *Colorado Law Review* 22:257.

Stryker, Robin, Danielle Docka-Filipek, and Pamela Wald. 2011. "Employment Discrimination Law and Industrial Psychology: Social Science as Social Authority and the Co-Production of Law and Science." *Law & Social Inquiry* 37(4):777–814.

Wheery, Robert J., and C. J. Bartlett. 1982. "The Control of Bias in Ratings: A Theory of Rating." *Personnel Psychology* 35:521.

White, Penny J. 2009. "Retention Elections in a Merit-Selection System: Balancing the Will of the Public with the Need for Judicial Independence and Accountability: Using Judicial Performance Evaluations to Supplement Inappropriate Voter Cues and Enhance Judicial Legitimacy." *Missouri Law Review* 74:635.

Wood, Rebecca, and Sylvia R. Lazos. 2009. *Reflections in Response to the Nevada Judicial Education Pilot Project*. UNLV William S. Boyd School of Law Legal Studies Research Paper No. 10-36. Las Vegas: University of Nevada Las Vegas. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1650764