4-5-2010

Sacrificing Diversity for "Quality": How Judicial Performance Evaluations are Failing Women & Minorities

Rebecca Wood
University of Nevada - Las Vegas

Sylvia R. Lazos
University of Nevada, Las Vegas -- William S. Boyd School of Law

Mallory Waters
University of Nevada - Las Vegas

Sacrificing Diversity for "Quality":
How Judicial Performance Evaluations are Failing Women & Minorities[1]

Rebecca Wood, Assistant Professor, Department of Political Science, University of Nevada Las Vegas

Sylvia R. Lazos, Justice Myron Leavitt Professor, William S. Boyd School of Law, University of Nevada Las Vegas

Mallory Waters, M.A. Candidate, Department of Political Science, University of Nevada Las Vegas

April 5, 2010

ABSTRACT: Because voters rely on judicial performance evaluations when casting their ballots, it is important that policymakers work diligently to compile valid, reliable and unbiased information about our sitting judges. This paper analyzes attorney surveys of judicial performance in Nevada from 1998-2008. The survey instrument is similar to those used throughout the country for judicial evaluation programs. Unfortunately, none of the readily-obtainable objective measures of judicial performance can explain away difference in scores based on race and sex. Minority judges and female judges score consistently and significantly lower than do their white male counterparts, all other things equal. These results are troubling, but consistent with the hypothesis that judicial performance evaluation surveys carry with them unexamined and unconscious gender/race biases.

I. Introduction

> *Unfortunately, judicial candidates, unlike eggs, do not come with a generally-recognized quality grade stamped on their forehead, visible to all.*
>
> —Jeffrey Jackson (2007, 126)

Clark County has a reputation for being one of America's most preeminent "judicial hell-holes" (ATRF 2008). The Elizabeth Halverson saga is just the latest embarrassing tale of judicial

[1] Paper prepared for presentation at the 2010 annual meeting of the Western Political Science Association, San Francisco, CA, April 1-3.

incompetence that threatens the independence of Nevada's judiciary. With the passage of Senate Joint Resolution 2 (SJR-2),[2] however, the state of Nevada is contemplating a complete overhaul of the way in which judges are selected and retained. As part of this proposed transition, a comprehensive system of state-sponsored judicial performance evaluations (JPEs) would be created.[3] While the proposed system is similar to JPE systems already in place around the country, we have found compelling evidence to suggest that the survey instrument in the proposed JPE system is likely to suffer from significant gender and race bias.

The appraisal of judicial performance, like other types of performance evaluations, "is fraught with many potential pitfalls and problems" (Kearney 1999, 483). Indeed, JPE programs "need to be designed and administered in a way that does not inadvertently harm the principles they are intended to promote" (Esterling 1998, 207). This is not an easy task, and many state programs have struggled through much trial and error while attempting to create fair and workable programs (Pelander 1998). Evidence about the validity, reliability and fairness of judicial evaluation programs is notably scarce (Kearney 1999), especially given the large number of such systems in use in the American states.

State JPE programs function in the pursuit of two goals: to provide information to the judges for the purposes of self-improvement, and to provide information to the public about the performance of their judges. Because voters rely on JPEs when casting their ballots (Esterling 1998), it is important that policymakers work diligently to compile valid, reliable and unbiased information about our sitting judges. The main instrument used in the most JPE programs is the survey. There are several important considerations that we must address in order for the surveys to be fair, useful and cost-effective. Among these is the potential for gender and racial bias.

II. Assessing Judicial Performance

The assessment of judges is often undertaken as an attempt to maintain a sense of accountability, and to prevent the judicial branch from falling victim to the low confidence ratings that are common in the executive and legislative bureaucracies. Judicial performance evaluation is nothing new; polls of attorneys seeking opinions on judicial performance date as far back as the 1870s (Feeney 1987). In large part, the more recent movement toward state-mandated judicial performance reviews began with in tandem with the shift toward merit plan

[2] Senate Joint Resolution No. 2 of the 74th Session (SJR-2). 2009, "Proposing to Amend the Nevada Constitution to provide for the initial appointment by the Governor of justices and judges and any subsequent retention of those justices and judges by election." Full text available at http://www.leg.state.nv.us/75th2009/Bills/SJR/SJR2_74_EN.pdf

[3] The proposed JPE program can be found in Thomas, Rebecca M., James T. Richardson, and Matthew Leone. "Nevada Judicial Evaluation Pilot Project: Final Report." Reno, NV: Grant Sawyer Center for Justice Studies, University of Nevada, Reno, April 28, 2009. (hereafter the *NJEPP Report*) and Thomas, Rebecca M., Victoria Springer, James T. Richardson and Matthew C. Leone. "Nevada Judicial Evaluation Project: Supplement to the Final Report," Reno, NV: Grant Sawyer Center for Justice Studies, University of Nevada, Reno, September 2, 2009 (hereafter the *NJEP Supplement*).

judicial selection. The state-run evaluations arose as a reaction to the dearth of candidate information available to voters in retention elections (Pelander 1998).

In 1985, in anticipation of the growing need for accountability and independence of the justice system, and with the birth of the first judicial performance evaluation program in Alaska nearly a decade old, the American Bar Association's (ABA) Special Committee on the Evaluation of Judicial Performance published guidelines intended to direct states toward the 'best practices' in the implementation of JPEs (American Bar Association 1985). Although the primary goal was to increase the ability of judges to improve themselves through constructive criticism and feedback, a secondary goal of the ABA report was to provide voters with valuable information on the quality of judicial candidates who were facing retention elections. Essentially, the ABA attempted to provide a roadmap for states seeking to create JPE programs by highlighting key criteria on which judges could adequately and appropriately be evaluated.

Since the issuance of the ABA's guidelines, 23 states have created or are currently in the process of developing JPE programs for the purposes of voter information and awareness, judicial self-improvement and judicial education (Rottman 2004). Among other things, the ABA report laid out five main categories of criteria on which judges should be evaluated: legal ability, integrity/impartiality, communication skills, professionalism/temperament and administrative capacity. The first column of Table 1 presents the subcategories from the ABA report.

The centerpiece of most state attempts to assess judges on these evaluation criteria is the survey (Pelander 1998). States go about the survey process in different ways, to be sure. Often, these evaluations are state-sponsored, although private and bar-run evaluation programs, like the *Las Vegas Review Journal's* "Judging the Judges" poll,[4] are found in several states. Although implications regarding legitimacy and reliability are undoubtedly different between state-operated JPE programs and those that are sponsored by private interests (Pelander 1998), the methodological choices and tools utilized in both categories are often very similar. State-sponsored programs tend to include jurors, litigants and/or other constituent groups in their survey mechanisms, while bar polls do not. Evaluation commissioners may rely on these respondent groups less, however, as commissioners report that lawyers provide the most helpful information (Esterling and Sampson 1998).

In 1998, the *Las Vegas Review-Journal* (*LVRJ*) began conducting judicial performance evaluations. These evaluations, which solicit survey responses from local attorneys, have been one of the important sources of information to Clark County voters in the intervening years. The evaluation of judicial performance is an important tool for protecting the quality and accountability of our judges. It is important, indeed, but it is no simple task. What is proposed in SJR-2, and in the *NJEPP Report*, is a comprehensive, state-sponsored system of judicial

[4] http://www.lvrj.com/hottopics/in_depth/judges/judgingthejudges2008.html. .

performance evaluations. The proposed plan shares many survey questions in common with the *RJ* poll.[5]

Among the states that have adopted some form of JPE program, nearly all utilize a survey as the key component of judicial quality determination. These surveys exhibit slight variations in terms of the groups to which they are distributed, the number of questions and the overall contribution to the overall judicial evaluation process. However, all hew closely to the ABA guidelines of 1985 in categorical references on the surveys as well as the survey questions themselves. Table 1 presents a comparison of the ABA guidelines, the *Review-Journal's* poll of attorneys and the attorney surveys used in other states (including the *NJEPP* proposal). The greatest variability among these is in the length of the various surveys. Although many states use surveys that are significantly longer than the *Review-Journal* poll, (e.g., Colorado's survey of district attorneys, which is composed of 36 questions) there is no evidence that longer surveys are buffers against racial and gender bias (Burger 2007). This is supported by the fact that, while longer, Colorado's survey does not present new questions or categories in terms of concepts. Essentially, Colorado is not asking new questions, but simply fracturing existing ones (Colorado Office of Judicial Performance Evaluation 2008).

Despite variations in survey length, the instruments in Table 1 show remarkable similarities. Currently, there is not much information available detailing the scores each individual judge received on the various survey questions. This analysis uses the *Review-Journal* poll, for which individual-level data are available. The similarity between the *Review-Journal* survey instrument and those used in other states implies that any bias problems identified in the *Review-Journal* poll may also be present in surveys that bear close resemblance to this instrument.

While many states rely heavily on surveys to assess judicial performance, "due to the dangers of using them without professional guidance and statistical controls, they should not be the sole basis for evaluation" (Brody 2000, 341). Scholars (Kearney 1999) and judges (IAALS 2008) agree that survey data cannot stand alone as a determinant of judicial quality. Although survey methods are often the primary tools, other measures such as discipline records, caseload evaluations, reversal rates, recusal rates and completion of continuing judicial education requirements are also important (American Bar Association 2005; Anderson 2000).

While most state JPE systems mention the importance of objective information, few include a systematic approach to integrating these data alongside the survey data. Even though states make haphazard use of this information at best, there is reason to believe that this information is important. Commissioners report that survey responses about legal knowledge are among

[5] The proposed state-sponsored survey instrument is a revision of the *LVRJ* bar poll, and it includes more questions targeted at rating judges on the same performance criteria (with the exception of the Public Outreach category, which was not included in the *LVRJ* bar poll). These revisions are similar in nature to those undertaken in Missouri in 2007. These changes actually *increased* mean differences in scores between male and female judges, as well as between white and black judges, yielding even lower scores for women and minority judges (Burger 2007).

the least helpful pieces of information they receive (Esterling and Sampson 1998). Supplementing these responses with objective measures is critical, as legal ability is a crucial part of what it means to be a "good" judge.

III. Gender and Race Bias in Judicial Evaluations

Gender and racial bias is systemic and pervasive. Judges subject to JPEs have long expressed concern about the presence of such bias in evaluation systems (Griffin 1994; IAALS 2008), but research is only now uncovering the answers to the critical question of how to minimize gender and racial bias in the judicial evaluation process (Durham 2000). This problem is certainly not unique to Nevada; of the twenty-two states that have state-sponsored JPE systems in place, not one has undertaken the work of systematically reviewing judicial performance surveys for evidence of systematic gender and race bias. Indeed, "there is a self-evident need to acquire more factual data about the operation of gender in the judicial performance evaluation process" (Durham 2000, 16). The same must be said for racial bias.

The most public debate about the fairness of judicial performance evaluations occurred in Missouri in 2007 when revisions were made to the evaluation survey administered by the Missouri Bar. These revisions increased mean differences in scores between male and female judges, as well as between white and black judges. At the request of the Mound City Bar, Dr. Burger (2007) undertook an analysis of Missouri's judicial performance evaluations. He found that gender- and race-based differences were statistically significant. Of the four groups that could be formed by classifying judges by both gender and ethnicity, female African-American judges were rated significantly lower than white female judges. Female judges, as a whole, ranked much lower than male judges (Burger 2007).

Within the body of anti-discrimination law, the powerful effect of unconscious biases in the evaluation process has long been recognized.[6] Moreover, there is a rich body of literature that has examined the effects of race and gender on the evaluation process in the context of employment decisions and teaching evaluations. Social psychologists have examined how implicit biases—meaning those that are automatic or unconscious, as distinct from explicit prejudices—diverge from self-reports and impact important decisions and social interactions. Negative attitudes and stereotypes, particularly with reference to race and gender, appear to have significant effects, even when people are unaware of the prejudice—and, in many instances, openly disavow it (Lane et al. 2007).

Another set of studies suggests that our brain processes facts through familiar scripts, which tend to emphasize existing subconscious ideas of gender or racial stereotypes (Harris and Carbado 2006). A meta-analysis conducted by the National Academy of Sciences indicates that women's performance ratings in jobs that are sex-typed as male (e.g., police officers, judges)

[6] See, for example, *Price Waterhouse v. Hopkins*, 490 U.S. 288 (1989) (finding that the evaluation of a female candidate for partner in an accounting firm was likely inflected by unconscious bias).

suffer when evaluated in comparison with men's ratings (USNAS 2006). Evaluations most likely trigger the evaluator's unexamined set of ideas linking gender and race to expectations of how these jobs should be performed. If women conform to gender role expectations in their jobs, they are rewarded. Women in male-typed jobs (i.e., jobs that carry authority, like judging) are not conforming to gender role expectations (Valian 1998). This literature predicts that women judges, because they work in a job that is sex-typed as male, will be evaluated more harshly than male judges.

Stereotypes can influence an evaluator's understanding of the meaning of a trait. Gender stereotypes shift baseline expectations for traits, as well as perceptions of what those traits entail (Sprague and Massoni 2005; Stathan et al. 1991). So, for example, in the area of student evaluations of teachers, students expect women to engage in a particular, feminine set of behaviors to satisfy a particular standard of being "approachable," while they have a separate set of expectations of what being "approachable" means for a male professor. In sum, a ten is not always a ten because students draw on different baselines depending on the professor's gender.

In the area of judicial performance evaluations, 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 gender (Durham 2000) and possibly race bias. In the 2008 *Review-Journal* Judging the Judges survey, male judges scored far higher than female judges on the question measuring "courtesy" (Sprague and Massoni 2005). The highest scoring female judge scored 63%, which was lower than the two *lowest* scoring male judges. One factor in this lopsided rating, as the *Review-Journal* conceded, is that two-thirds of the attorneys surveyed were men (Geary 2008).

IV. Data and Measures

The dependent variable for our analysis comes from data from the *LVRJ*'s "Judging the Judges" biennial survey of local attorneys rating Clark County judges and justices.[7] This survey was administered by a private research firm, and the results were made public—minus the attorney's comments. The survey was originally co-administered by the *Las Vegas Review-Journal* and the Clark County Bar but, starting in 2006, has been administered by the *Review-Journal* alone (Whitely 2006). In 2008, the survey covered 68 Clark County judges and the justices sitting on the Nevada Supreme Court. About 800 area lawyers responded to the survey after averring to having had direct experience dealing with the sitting judges (Hopkins 2008).

[7] The "Judging the Judges" judicial performance evaluation survey was originally co-administered by the *Las Vegas Review-Journal* and the Clark County Bar but, starting in 2006, has been administered by the *Review-Journal* alone (Whitely 2006)

Our data include the universe of judicial evaluations from these biennial surveys from 1998-2008.[8] The survey includes the questions outlined in the second column of Table 1.[9] These questions address most of the categories recommended in the ABA's guidelines (American Bar Association 1985). The last question in the survey reads, "Taking everything into account, would you recommend retaining this judge on the bench?" The percentage of respondents answering "yes" to this question represents the retention score. This is the dependent variable in our analysis.[10]

The independent variables of interest in this analysis are gender and minority status. A dummy variable is included to indicate whether the judge is female (1) or male (0), and another indicates whether the judge is of a minority race (1) or not (0). The retention score exhibits a consistent, statistically significant difference among judges based on both gender and race. Figure 1 shows the frequency of votes in favor of retention broken down by gender. The graph demonstrates that the mean for male judges is significantly higher[11]—more than 10.5 points higher—than for female judges. The distribution of scores for male judges is skewed, with more high scores than would be predicted by the normal distribution. We have found even more significant differences between the scores of racial/ethnic minority judges and their white counterparts.[12]

Of course, this difference may be a result of the "affirmative action" hypothesis. If female and minority judges, on average, actually are less qualified and less able on the bench, then the

---

[8] All of the relevant reports, with the exception of the 1998 survey, can be found on the *Las Vegas Review-Journal* website: http://www.lvrj.com/hottopics/in_depth/judges/judgingthejudges2008.html. This clearinghouse of "Judging the Judges" data also includes links to the survey methodology description as well as the *LVRJ*'s coverage of the survey results over the years 2000-2008.

[9] Between the 2000 and 2002 survey, the questionnaire was shortened from 15 questions to 13 questions. The new, shorter version did not omit questions, but instead collapsed some of the questions.

[10] In some of the survey reports, an "adequacy" score was also provided. The questions on the survey, with the exception of the retention question, asked attorneys to rate judges on a three-point scale: not adequate, adequate or more than adequate. The adequacy score was calculated by taking the percentage of "adequate" or "more than adequate" ratings the judges received. We ran our model with this adequacy score as the dependent variable, and the results were quite similar to what is presented here.

[11] Independent Samples Difference of Means test for weighted data, such that the scores for each judge are averaged across the available survey years, assuming unequal variances (Levene's Test for Equality of Variances F=76.601, p≤.001, confirmed by an Independent-Samples Mann-Whitney Test (p≤.001) shows a statistically significant difference between the mean for women (66.93) and for men (77.63) (t=-29.773, p≤.001), confirmed with an Independent-Samples Difference of Medians Test (p≤.001) . The unweighted data, in which each survey score is counted equally regardless whether the judge appears more than once in the dataset yields similar results with a similar magnitude of mean difference between male and female judges.

[12] Independent Samples Difference of Means test for weighted data, such that the scores for each judge are averaged across the available survey years, assuming unequal variances (Levene's Test for Equality of Variances F=6.047, p≤.01, confirmed by an Independent-Samples Mann-Whitney Test (p≤.001) shows a statistically significant difference between the mean for minority judges (61.93) and for white judges (74.54) (t=-19.094, p≤.001), confirmed with an Independent-Samples Difference of Medians Test (p≤.001) . The unweighted data, in which each survey score is counted equally regardless whether the judge appears more than once in the dataset yields similar results with a similar magnitude of mean difference between minority and white judges.

difference in scores would represent a true difference in judicial performance. In other words, the observed differences could be attributed not to bias, but to real differences in judicial quality. In order to control for this possibility, and other possible explanations for variation in retention scores, a number of control variables are included in the model.

The control variables are intended to serve as proxy measures for the quality of the judges in the study. The most obvious source of objective information about legal ability is the reversal rate (Brody 2000; Posner 2000). When a judge has a low reversal rate, this is an indication that the judge is interpreting the law correctly in the eyes of the appellate court (Feeney 1987)
For this study, reversal rate figures were derived from a Lexis-Nexis search for each judge spanning the years 1998-2009. Because our sample includes judges from across the spectrum of Nevada's jurisdictions, the measure of reversal rate is calculated as the distance between an individual judge's reversal rate and the mean for his or her court. The mean value of this variable is zero, and this represents the mean reversal rate for the level of court. In our dataset, the reversal rate distance ranges from -0.35 (which is a lower or "better" reversal rate than the court mean) to 0.13.

Certainly, the reversal rate has the potential to contain artifacts of race and/or gender bias. As Walker and Barrow observe, "[i]f the decisions of female and black judges are of a low quality or if practicing attorneys perceive them to be vulnerable to reversal on appeal, then a pattern of relatively high appeals rates for nontraditional judges should occur" (1985, 612). To investigate this possibility, difference of means tests were conducted on the both the reversal rate distance and the appeal rate distance. The appeal rate distance is calculated the same way as the reversal rate distance, but using the number of times a judge was appealed as compared with the mean of the relevant court. Because this measure necessarily uses a raw number as opposed to a rate, it is weighted by the number of evaluation years we have for each judge. For our data, appeal rate ranges from -62.6 to 171.9. Difference of means tests reveal no significant bias effects in either of these variables.[13] This is in keeping with findings from previous research on diversity on the federal bench (Walker and Barrow 1985).

Another proxy measure for judicial quality is whether or not the judge was first appointed to bench. In Nevada, unlike most other states with competitive judicial elections, judicial vacancies are filled by via a merit selection plan of sorts (Flango and Ducat 1979). This system is enshrined in Article VI of the Nevada Constitution, and it requires the Commission on Judicial Selection to devise a list of nominees from which the governor chooses one appointee. Having been appointed at first instance does appear to confer some benefits on judges; most notably, these judges obtain incumbency status without first having to stand for election (Glick 1978).

There is mixed evidence at best to suggest that judges selected through the merit plan are of higher quality than elected judges (Bonneau and Hall 2009). Even so, there is reason to believe

[13]The results of the difference of means tests for female judges (on reversal rate, F=0.487; on appeal rate, F=.515) and for minority judges (on reversal rate, F=1.003; on appeal rate, F=0.089) yielded no statistically significant differences in either reversal rates or appeal rates.

that attorney respondents in the *LVRJ* survey may rate these appointed judges more highly. First, attorneys tend to support the merit plan because the state bar plays such a critical role in the selection process. In addition, these judges have had the luxury of serving on the bench for a while without having faced the public criticism that elections tend to bring. Additionally, because Nevada's current selection system is nonpartisan, elected judges face no vetting process prior to taking their place on the ballot, with the exception of having to collect signatures on a petition. The absence of a vetting process (which, in partisan elections, is provided through the political party system) may lead to lower-quality candidates.

The nature of a judge's legal education is another proxy for judicial performance, "though like most proxies a rough one" (Posner 2005). The prestige of a judge's legal education carries with it certain implications for the judge's perceived level of legal ability (Glick 1978).[14] Research on the American Bar Associations ratings of federal judicial nominees indicate that law school prestige is an important covariate of quality scores (Slotnick 1983). For this analysis, a variable representing the current tier ranking of the judge's law school education as per the 2010 rakings from U.S. News & World Reports.[15] Alumni information was gathered from various online sources, including official judge biographical blurbs, resumes and newspaper articles. Of the 94 judges in this dataset, 36 came from the very best schools and 33 came from second-tier schools. On the lower prestige end, six judges came from tier three schools, sixteen from tier four schools, and three did not graduate from law school at all.

Previous judicial experience is a time-tested measure of judicial decision making (Epstein et al. 2003). In the realm of judicial quality, it stands to reason that a judge who has served longer has more opportunity to learn and master the trade (Haire 2001). Research on federal court judges suggests that newer judges experience acclimation effects as they familiarize themselves with the job of judging (Brenner and Hagle 1996; Hettinger et al. 2003). Of course, it is possible that the longer one stays on the bench, the more likely it is that the respondent attorneys will see them as "old-timers" or otherwise past their prime. This is certainly part the crux of the argument for mandatory retirement ages. As with the information about law schools, judicial experience information was gathered from publicly available sources. Judges in our sample ranged from zero to thirty years on the bench.

Disciplinary complaints are a particularly important signal about a judge's performance (Posner 2005). To be sure, some lawyers may be unwilling to file formal complaints against judges for fear of future retribution (Jackson 2007). Even still, disciplinary records are the single most important objective measure of judicial integrity available to researchers. It is not costless to file a disciplinary complaint against a judge, but it is relatively easy. For this reason, we have

[14] Glick (1978) also suggests that the in-state/out-of-state law school distinction is important. In Nevada, this distinction is less helpful. Nevada's first law school, the William S. Boyd School of Law at the University of Nevada Las Vegas, did not graduate its first class until 2001.

[15] These rankings can be found online at http://grad-schools.usnews.rankingsandreviews.com.

elected to code our disciplinary record variable in terms of the outcome of the disciplinary complaint. These outcomes were ranked on an ordinal scale of zero to eight.[16]

Not all scandals rise to the level of disciplinary complaints, though. Unfortunately, even the mere suggestion of impropriety may drive attorney opinion against the judge who is the subject of public scandal. The *Review-Journal* bar poll asks respondents to evaluate only those judges they have appeared before in the evaluation period, but judges report that they frequently receive more evaluations than the number of attorneys appearing in their courtrooms. All respondents may be swayed by a judge's connection to public scandal as, at the very least, unbecoming of a member of the judiciary. The ratings of those attorneys without recent, first-hand experience with the judge may be particularly susceptible to change based on this sort of information. For this reason, we have included a dummy variable to flag those judges who have been publicly connected—rightly or wrongly—to scandal. This information comes from publicly available news sources, mostly from the pages of the *Review-Journal* itself. A full 23% of the observations in this sample include a reported scandal.

Another important consideration is the relative weight we give to the various judge ratings of the *Review-Journal* poll. While the response rates of this poll often fall below the generally-accepted level of 50% of surveys sent out (Brody 2000), the assessment of response rate is actually quite a bit more complicated than it seems. In the *Review-Journal* methodology, like many of the other state surveys, the respondents are instructed to answer questions only if they have professional experience with the judges. In 2008, for example, researchers sent out 4,237 survey invitations, but only 799 attorneys responded (Downey 2008). This yields an incredibly low (and suspect) response rate of judge 18.9%. Beyond this, none of these attorneys had experience with all of the 68 judges evaluated in the 2008 survey. For the 2008 results, an average of 202 attorneys evaluated each judge, with a range of 44 to 387. Indeed, the results calculated with the input of very few attorneys will be less reliable than the others. As such, a weighted model of evaluation scores is estimated.[17]

Because we have unbalanced panel data, estimate a pooled weighted least squares (WLS) model with Driscoll and Kraay (1998) standard errors. This adaptation of the Beck and Katz (1995) approach allows for the regression to be adapted based on a robust estimate of the error structure of the model (Hoechle 2007). As we have several biannual estimates for most judges, we cannot assume that these observations are independent. This model allows the standard errors to be calculated with this systematic dependence in mind.

[16] The scale for this ordinal variable is as follows: 0=no complaint, 1=complaint dismissed, 2=required course, 3=required course & public apology, 4=public reprimand, 5=public reprimand & fine, 6=censure, required course and fine, 7=removal from bench, 8=removal from bench and permanently barred from holding public office in Nevada.

[17] The model was also estimated without this weight and using the Beck and Katz (1995) panel corrected standard errors method. The results of all of these analyses were similar to what is reported here.

V. Evidence of Bias in the *Review-Journal* Survey Instrument

The results of our analysis are presented in Table 2. Unfortunately, none of the readily-obtainable objective measures of judicial performance can explain away difference in scores based on race and sex. Minority judges and female judges score consistently and significantly lower than do their white male counterparts, all other things equal. These results are consistent with the hypothesis that judicial performance evaluation surveys carry with them unexamined and unconscious gender/race biases (Durham 2000).

In the full model, only two of the control variables fail to help explain variation in retention recommendation rates. The prestige of law school education is insignificant, as is the reversal rate distance from the court mean. Both of these findings are somewhat troubling, and they show that the attorney ratings do not tend to take these important measures of legal ability into account. Part of the reason may be that, as compared with measures of integrity, legal knowledge questions make up a small fraction of the questions on the *LVRJ* survey.

The rest of the variables are statistically significant in the full model, and are included in the reduced model on the right-hand side of Table 2. The intercept for this model is about 86 points, which is the starting place for judges when all other variables are held constant at zero. Judges who are appointed at first instance get a two-point boost in scores. Substantively, this is not very big, and it may reflect the legal profession's general preference for an appointment system of judicial selection.

Years of experience are also significant and negative, indicating that for each additional year of experience on the bench, the judge loses more than a third of a point. This indicates that lawyer respondents tend to rate judges with longer tenure more harshly.[18] The integrity measures are also statistically significant. As a judge's disciplinary result moves one step up the scale of severity, the judge can expect to lose about two points. When the judge is connected with a public scandal, he or she can expect a loss of six points.

Of course, the independent variables of interest are both statistically significant and of very high magnitude. When all of these control variables are equal, female judges can expect to score 11.5 points lower than their male colleagues. Minority judges score nearly 14 points lower than non-minority judges. Even after controlling for the affirmative action hypothesis, these judge demographic variables still account for the bulk of the substantive variation in retention scores.

[18] A nonlinear version of this relationship was estimated, operating on the premise that experience may have a diminishing returns effect over time. In other words, we tested the hypothesis that some experience would increase scores, but judges who had been on the bench for decades would be penalized as being "too old" or "out of touch." We found no evidence to support this hypothesis.

VI. Discussion

Because we have controlled for objective measures of judicial quality, the "affirmative action" hypothesis is no longer a convincing explanation for the systematically lower retention scores of female and minority judges. Even when these judges are on par with their peers in terms of experience, education and integrity, their scores are still drastically lower than their white male counterparts. Of course, it is very difficult to arrive at these objective measures, and those used in this analysis are admittedly imperfect. However, each of these judge quality control variables *ought* to be related to an assessment of a judge's performance. If the affirmative action hypothesis were true, we would expect these variables to explain much of the variance in retention scores. Unfortunately, this is not the case.

Indeed, some of the more theoretically important variables in the analysis, legal education and reversal rates, are insignificant in the model. This finding points to a major weakness in the existing JPE survey construction. The null finding for the reversal rate distance variable is particularly disheartening. Judges who make more "correct" decisions are not rewarded for it in the retention scores. Reversal rates are among the most relevant and quantifiable objective measures we have, and they have no effect on the retention scores which, in turn, greatly affect the choices made by the voting public (Esterling 1998).

In the area of judicial performance evaluations, 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 gender (Durham 2000) and possibly race bias. In the 2008 *Review-Journal* Judging the Judges survey, male judges scored far higher than female judges on the question measuring "courtesy" (Sprague and Massoni 2005). The highest scoring female judge scored 63%, which was lower than the two *lowest* scoring male judges. One factor in this lopsided rating, as the *Review-Journal* conceded, is that two-thirds of the attorneys surveyed were men (Geary 2008).

There are reasons to think that these results are not aberrational. Women judges at every level report that their authority in the courtroom is challenged in ways that do not mirror what happens to their male colleagues. Should they fail to take control of their courtroom, they are perceived as weak and indecisive—or even incompetent. If they assert their authority, though, they will be labeled as unduly punitive. In either case, their evaluations by members of the Bar—who, under the current system, are the primary source of performance assessments—are negatively impacted. These concerns are heightened further for minority women judges, who have faced even more intense scrutiny and challenges to their competence.[19] Because the

[19] In the Mound City Bar study conducted by Dr. Burger (2007), African American female judges were rated almost 15 points lower than white male judges, on the average. In other areas of legal world, such as private law firms, African American women also suffer harsh scrutiny and discrimination. *See* American Bar Association, *Visible Invisibility: Women of Color in Law Firms* (2006). The study cites data from the National Association of Law Placement showing 81 percent of minority female associates left their jobs within five years of being hired. Close to half of the respondents were able to cite specific instances of conscious and

survey instruments used across the country bear such strong resemblance to the "Judging the Judges" survey (see Table 1), these findings raise significant questions about the ability of JPEs to promote and protect the diversity of our country's state benches.

These insights have important implications for assessing the evaluation process. As there are increasing calls for reliance on JPEs as a way of ensuring quality standards, it is imperative that these processes not reproduce—even inadvertently—a system that disfavors groups like women and minorities, who have been historically underrepresented in the judiciary. Unfair and biased evaluations do not only harm the individuals subject to them, but would have far-reaching and deleterious effects on the courts as an institution. Voters who have access to state-sponsored judicial performance evaluations rely on that information when casting their votes (Esterling 1998). For this reason, it is fundamentally important for the democratic process of judicial elections that JPEs refrain from providing biased, uneven or unfair coverage of judicial candidates.

unconscious discrimination. We have not tested this interaction term here as, unfortunately, minority women judges are nearly a null set in our data.

Figure 1 -- "Yes" Votes for Retention in *LVRJ* Bar Poll Survey by Gender, 1998-2008

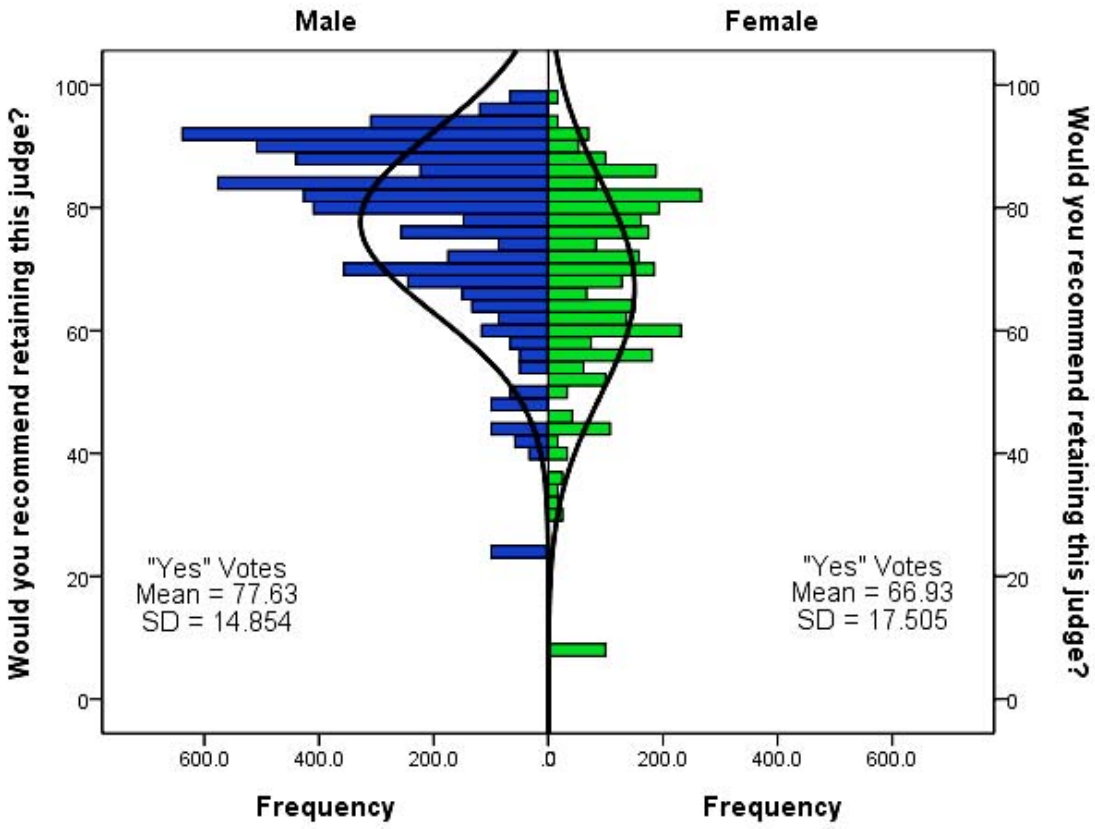

Table 1 -- Comparing Questions Between the APA Guidelines, the *LVRJ* Survey, and Other State Surveys

ABA Guidelines for JPEs	The *Las Vegas Review-Journal* "Judging the Judges"	Use by Other States
Legal Ability (5-1)		
1-1: Legal reasoning ability.	The judge's rulings, whether regarding civil issues, criminal sentencing, or contempt are appropriate.	AZ, CO, HI, KS, NM, NC, NV, WA
1-2: Knowledge of substantive law.	The judge demonstrates familiarity with the case record and documents, and fairly weighs all evidence and arguments before rendering a decision.	AZ, CO, HI, KS, MO, NM, NC, NV
1-3: Knowledge of rules of procedure and evidence.	The judge properly applies the law, rules of procedure, and rules of evidence.	AK, AZ, CO, KS, MO, NM, NC, NV, WA,
1-4: Keeping current on developments in law, procedure, and evidence.	N/A	NC, NV
Integrity and Impartiality (5-2)		
2-1: Avoidance of impropriety and the appearance of impropriety.	The judge's professional conduct is free from impropriety and the appearance of impropriety.	AK, AZ, HI, KS, MO, NM, NC, NV, WA
2-2: Treating all people with dignity and respect.	N/A	AK, AZ, CO, HI, MO, NC, NV, WA
2-3: Absence of favor or disfavor toward anyone, including but not limited to favor or disfavor based upon race, sex, religion, national origin, disability, age, sexual orientation, or socioeconomic status.	The judge's conduct is free from bias on the basis of race, ethnic origin, gender, and religion (posited in several individual questions).	AK, AZ, CO, HI, KS, MO, NM, NC, NV, WA
2-4: Acting fairly by giving people individual consideration.	N/A	CO, KS, MO, NV
2-5: Consideration of both sides of an argument before rendering a decision.	The judge demonstrates familiarity with the case record and documents, and fairly weighs all evidence and arguments before rendering a decision.	AK, CO, KS, MO, NV
2-6: Basing decisions on the law and the facts without regard to the identity of the parties or counsel, and with an open mind in considering all issues.	The judge's conduct is free from bias on the basis of parties or attorneys involved in the action.	AK, CO, HI, KS, NC, NV
2-7: Ability to make difficult or unpopular decisions.	N/A	AK, CO, KS, MO
Communication Skills (5-3)		
3-1: Clear and logical oral communication while in court.	The judge clearly explains the basis for his or her decisions.	AZ, CO, HI, KS, MO, NM, NC, NV, WA
3-2: Clear and logical written decisions.	The judge clearly explains the basis for his or her decisions	AZ, CO, HI, KS, MO, NM, NC, NV, WA

ABA Guidelines for JPEs	**The *Las Vegas Review-Journal* "Judging the Judges"**	**Use by Other States**
Professionalism & Temperament (5-4)		
4-1: Acting in a dignified manner.	N/A	AZ, CO, HI, KS, MO, NC, NV, WA
4-2: Treating people with courtesy.	The judge is courteous.	AK, AZ, CO, KS, MO, NC, NV, WA
4-3: Acting with patience and self control.	N/A	AZ, KS, NC, WA
4-4: Dealing with pro se litigants and litigation fairly and effectively.	N/A	CO, NC
4-5: Participating and providing leadership to an appropriate degree in professional development activities and in jurisdiction- wide and statewide court improvement and judicial education activities.	N/A	NV
4-6: Promoting public understanding of and confidence in the courts.	N/A	AZ, NV
Administrative Capacity (5-5)		
5-1: Punctuality and preparation for court.	The judge is punctual in convening court, keeps business moving, and does an amount of work fair to taxpayers and other judges.	AK, CO, HI, KS, MO, NM, NC, NV, WA
5-2: Maintaining control over the courtroom.	N/A	AK, CO, KS, MO, NM, NC, NV, WA
5-3: Appropriate enforcement of court rules, orders, and deadlines.	N/A	CO, KS, NM, NC, NV, WA
5-4: Making decisions and rulings in a prompt, timely manner.	The judge issues orders, judgments, decrees, or opinions without unnecessary delay.	AK, AZ, CO, KS, NM, NC, NV, WA
5-5: Managing his or her calendar efficiently.	N/A	CO, KS, NM, NC, NV,WA
5-6: Using settlement conferences and alternative dispute resolution mechanisms as appropriate.	N/A	CO, KS, NV
5-7: Demonstrating appropriate innovation in using technology to improve the administration of justice.	N/A	None
5-8: Fostering a productive work environment with other judges and court staff.	N/A	None
5-9: Utilizing recruitment, hiring, and promotion policies and practices to ensure that a pool of qualified applications for court employment is broad and diverse.	N/A	None
5-10: Acting to ensure that disabilities and linguistic and cultural differences do not limit access to the justice system.	N/A	None

Table 2 – Pooled OLS Model * of *LVRJ* "Judging the Judges" Retention Scores

	Full Model			Reduced Model		
	Coef.	**Std. Err.**	**z-score**	**Coef.**	**Std. Err.**	**z-score**
Constant	89.24	(1.65)	53.03***	86.18	(1.08)	79.47***
Reversal Rate	-4.37	(4.45)	-0.98	- - -	- - -	- - -
First Appointed	1.76	(0.60)	2.95**	1.97	(0.57)	3.44***
Law School Tier	-0.56	(0.39)	-1.43	- - -	- - -	- - -
Years Experience	-0.39	(0.15)	-2.66**	-0.39	(0.15)	-2.66**
Discipline Result	-1.94	(0.91)	-2.14*	-1.97	(0.84)	-2.33*
Reported Scandal	-5.65	(0.62)	-9.18***	-6.03	(0.76)	-7.97***
Female Judge	-11.27	(3.66)	-3.08**	-11.54	(3.41)	-3.39***
Minority Judge	-14.27	(1.59)	-8.98***	-13.84	(1.87)	-7.41***
* Pooled Ordinary Least Squares Regression with Drisoll-Kraay (1998) Standard Errors	N(obs)=364; N(judges)=94 F (8,93) = 1980.93 *** Root Mean Squared Error = 316.73			N(obs)=364; N(judges)=94 F (6, 93) = 3536.76 *** Root Mean Squared Error = 318.16		

VII. Works Cited

American Bar Association. 1985. "Guidelines for the Evaluation of Judicial Performance." Washington, D.C.

———. 2005. "Guidelines for the Evaluation of Judicial Performance with Commentary." Chicago, IL.

Anderson, Seth S. 2000. "Judicial Retention Evaluation Programs." *Loyola L.A. Law Review* 34:1375.

ATRF. 2008. "Judicial Hellholes 2008/2009." Washington, D.C.: American Tort Reform Foundation.

Beck, Nathaniel, and Jonathan Katz. 1995. "What To Do (and Not To Do) With Time-Series Cross-Section Data." *American Political Science Review* 89 (3):634-47.

Bonneau, Chris W., and Melinda Gann Hall. 2009. *In Defense of Judicial Elections*. Edited by M. J. Streb. New York: Routledge.

Brenner, Saul, and Timothy M. Hagle. 1996. "Opinion Writing and Acclimation Effects." *Political Behavior* 18 (3):235-61.

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

Burger, Gary K. 2007. "Attorney's Ratings of Judges: 1998-2006." Mound City, MO: Report to the Mound City Bar.

Colorado Office of Judicial Performance Evaluation. *Judicial Performance Reviews* [Website] 2008 [cited March 17, 2010]. Available from http://www.coloradojudicialperformance.gov/review.cfm.

Downey, Nancy. 2008. "Las Vegas Review Journal Judicial Performance Evaluation 2008." Las Vegas, NV: Downey Research Associates.

Driscoll, John C., and Aart C. Kraay. 1998. "Consistent Covariance Matrix Estimation With Spatially Dependent Panel Data." *Review of Economics and Statistics* 80 (4):549-60.

Durham, Christine M. 2000. "Gender and Professional Identity: Unexplored Issues in Judicial Performance Evaluation." *Judges' Journal* 39 (2):13-6.

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. 1998. "Judicial Accountability the Right Way." *Judicature* 82:206.

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.

Flango, Victor Eugene, and Craig R. Ducat. 1979. "What Difference Does Method of Judicial Selection Make? Selection Procedures in State Courts of Last Resort." *Justice System Journal* 5 (25-45):25.

Geary, Frank. 2008. "Lawyers Rate Female Jurists as Less Courteous than Men." *Las Vegas Review-Journal*, May 21.

Glick, Henry R. 1978. "The Promise and the Performance of the Missouri Plan: Judicial Selection in the Fifty States." *University of Miami Law Review* 32 (3):509-43.

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.

Harris, Cheryl I., and Devon W. Carbado. 2006. "Loot or Find: Fact or Frame?" In *After the Storm: Black Intellectuals Explore the Meaning of Hurricane Katrina*, ed. D. Troutt. New York, NY: The New Press.

Hettinger, Virginia A., Stefanie A. Lindquist, and Wendy L. Martinek. 2003. "Acclimation Effects and Separate Opinion Writing on the United States Courts of Appeals." *Social Science Quarterly* 84 (4):792.

Hoechle, Daniel. 2007. "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence." *Stata Journal* 7 (3):281-312.

Hopkins, A. D. 2008. "Judging the Judges: Nearly 800 Lawyers Turn in Judicial Performance Evaluation Grades." *Las Vegas Review-Journal*, May 18.

IAALS. 2008. "The Bench Speaks on Judicial Performance Evaluation: A Survey of Colorado Judges." Denvery, CO: Institute for the Advancement of the American Legal System.

Jackson, Jeffrey. 2007. "Beyond Quality: First Principles in Judicial Selection and their Application to a Commission-Based Selection System." *Fordham Urban Law Journal* 34:125.

Kearney, Richard C. 1999. "Judicial Performance Evaluation in the States." *Public Administration Quarterly* 22:468.

Lane, Kristin A., Jerry Kang, and Mahzarin Banaji. 2007. "Implicit Social Cognition and the Law." *Annual Review of Law and Social Science* 19:1-25.

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

Posner, Richard A. 2000. "Is the Ninth Circuit Too Large? A Statistical Study of Judicial Quality." *The Journal of Legal Studies* 29 (2):711-9.

———. 2005. "Judicial Behavior and Performance: An Economic Approach." *Florida State University Law Review* 32:1259-79.

Slotnick, Elliot E. 1983. "The ABA Standing Committee on Federal Judiciary: A Contmporary Assessment--Part 1." *Judicature* 66:349.

Sprague, Joey, and Kelley Massoni. 2005. "Student Evaluations and Gendered Expectations: What we Can't Count Can Hurt Us." *Sex Roles* 53:779-93.

Stathan, Anne, Laurel Richardson, and Judity Cook. 1991. *Gender and University Teaching: A Negotiated Difference*. New York: SUNY Press.

USNAS. 2006. "Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering." Washington, D.C.: United States National Academy of Sciences.

Valian, Virginia. 1998. *Why So Slow? The Advancement of Women*. Boston, MA: MIT Press.

Walker, Thomas G., and Deborah J. Barrow. 1985. "The Diversification of the Federal Bench: Policy and Process Ramifications." *Journal of Politics* 47:596-617.

Whitely, Joan. 2006. "Judging the Judges: About the Survey." *Las Vegas Review-Journal*, Apr. 30.