Implicit Gender Bias in State-Sponsored Judicial Performance Evaluations:

A Preliminary Analysis of Colorado's JPE System, 2002-2012 [1]

Rebecca D. Gill
University of Nevada Las Vegas

Kenneth J. Retzl
University of Nevada Las Vegas

One of the important innovations of the merit system of judicial selection is its ability to maintain electoral accountability for judges while removing the need for judges to square off head-to-head in competitive elections. The retention elections used by most merit plan systems allow the citizens to remove underperforming judges at the ballot box, but are designed to minimize the need for campaigning and fundraising that competitive elections often require.

A side effect of the absence significant campaigning, however, is that the voter is left with little information on which to evaluate the judges who are standing for retention. In contested elections, the campaign process (and party labels where partisan elections are used) gives voters important information about the incumbent and the challenger (Bonneau and Hall 2009). The challenger in the race has an incentive to make public whatever shortcomings the incumbent judge might have. Challengers typically seek out weak incumbents, increasing the likelihood that voters will be exposed to information about the weak incumbent's record. In the absence of a motivated challenger, however, voters have virtually no information on which to base their decision whether to retain the incumbent judge. As such, we see low levels of voter turnout and high levels of ballot roll-off in these races.

States that use retention elections to provide accountability for their merit-appointed judges have recognized this problem. The most important thing they have done to counteract this dearth of information is the creation of formal, state-sponsored judicial performance evaluation (JPE) programs. These programs aspire to provide useful, fair, and relevant information for voters to use when making the decision whether to retain a judge. The American Bar Association has contributed to the process by publishing a set of guidelines and

best practices for the development of JPE instruments and institutions (American Bar Association 1985, 2005).

Although some have dismissed "the concerns about the subjectivity of legitimate evaluation factors" as "infinitesimal" (White 2001, 1075), even proponents of such systems have long recognized the importance of addressing concerns about the "fairness of survey methodologies and evaluation commission procedures" (Andersen 2000, 1376). The judicial evaluation system in Colorado grew out of a self-conscious attempt to do just that. The move toward a more complex, formalized JPE system in Colorado was spurred on by the work of several independent organizations, including the Colorado Judicial Institute. The bill that created the Judicial Performance Commissions was signed into law in 1988. Those who supported it at time saw it as a model of the ideal JPE system, which should create an avenue for the active involvement of the citizens in the evaluation of state judges (Mahoney 1989) .

One of the most worrisome sources of unfairness in JPEs comes from unconscious gender bias. 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. Ardent supporters of the merit plan and state-sponsored JPE systems have begun to address the possibility that the way JPEs are implemented in the states may be uniquely subject to implicit gender bias (Knowlton and Reddick 2012). We applaud efforts to improve the fairness of JPEs, and we envision this paper as contributing to this goal.

After a brief review of the literature on performance evaluation and implicit gender bias, we provide an introduction to the Colorado JPE system. We summarize a new dataset of information about the results of the Colorado JPEs from 2002-2012. Finally, we provide a preliminary analysis of the predictors of JPE decisions. We find that the outcome of the JPE process relies strongly on the attorney survey component. We also find a statistically significant difference between the success of male and female judges at both the Committee stage and the attorney survey stage.

IMPLICIT BIAS AND PERFORMANCE EVALUATION

Traditionally, gender and race discrimination have been understood to be products of conscious motive or intent (Krieger 1995). But it is likely that this kind of gender- and race-related hostility makes up only a fraction of the bias we might see in professional performance evaluations. For example, male legal professionals tend to perceive much less gender bias in the workplace than do their female colleagues (Coontz 1995). This is exemplified in the introductory quote. Research also shows that, even in the context of increasing diversity initiatives on the part of 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 lags, 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: unconscious bias. The process of simplifying and categorizing our environment, which exists is a necessary condition for most higher-level cognitive function, processes people just as it does letters, shapes, and

4

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 upon the accumulated stereotypes about a particular outgroup, resulting in inaccuracy and unfairness based on race or gender.

The social science evidence for unconscious 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 holds 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 usually not 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 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 unconscious bias. And this kind of bias happens much more furtively than bias based on explicit racism or sexism. Unconscious bias theory is a logical extension of implicit social cognition. 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 our evaluations and behaviors of others without our conscious consent.

A few of aspects of unconscious bias 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). This often results in a paradox or "double bind" for women in the legal profession because they are penalized in their performance evaluations both for being too masculine and for not fitting the masculine stereotype for the job (Bowman 1998).

A second important characteristic of unconscious bias is the fact that subjective evaluation criteria exacerbate discriminatory employment decisions (Fiske et al. 1991). 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?" in Nevada's Judging the Judges survey has 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 may also exacerbate unconscious gender and race bias. Evidence suggests that the anonymity of evaluations increases the effects of implicit bias (Hekman et al. 2010). 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 information. 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 is a male-stereotyped position. The types of questions asked are generally subjective. These are anonymous surveys. They are often done quickly, as attorneys are asked to rate several judges at a time on their performance over the past two years. In all, attorney surveys of judicial performance may be even more likely than other performance evaluations to suffer from unconscious gender and race bias.

These insights have important implications for assessing the evaluation process. There are increasing calls for reliance on JPEs as a way of ensuring quality standards in the judiciary (White 2009). In this context, it is imperative that JPEs 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 they can have far-reaching and deleterious effects on the courts as an institution.

Survey research is a very complicated task, but many JPE programs have not been put together by experts in assessment methodologies. Typically, committees are made up of attorneys who lack such expertise. These committees often engage a single consultant who may or may not have all of the necessary areas of expertise (Wood and Lazos 2009). To date, there has been no comprehensive assessment of the potential for unconscious race and gender bias in Colorado's JPE surveys. In the absence of this research, performance evaluation committees are forced to proceed blindly, hoping that the evaluations they are conducting do not systematically disadvantage women and minorities.

The limited evidence that we have so far indicates that this is a risky gamble. Most of the previous research underlying this cognitive bias theory has relied on self-reported feelings of bias; the research presented here provides a more systematic evaluation of gender and race based disparities in actual performance evaluations.

To date, the small amount of research that has been done on bias in judicial performance evaluations has focused on what is arguably the most subjective question on the survey: "Should Judge X be retained?" (Burger 2007; Gill et al. 2011). The stereotyping that leads to unconscious bias is exacerbated in situations where the evaluation criteria to be used are ambiguous. Certainly this is the case in the retention question. But JPEs around the country use more than just one yes/no question; all of them include a series of more specific questions intended to capture a particular dimension of judicial quality. The JPE programs currently in

existence rely heavily on the questions found in the ABA Guidelines (American Bar Association 1985).[2]

Colorado Judicial Performance Evaluation Process

The Evaluation Process

The governor of Colorado has the authority (responsibility) of appointing judges to the county, district, appellate, or supreme courts. By law, each judge serves an initial term of two years and must stand for retention during the next general election. If retained by the voters, the judge's term is dependent on the court over which they preside. County judges serve for four years, district judges serve for six years, court of appeals justices serve for eight years, and Supreme Court justices serve terms of ten years. After a judge's term is expired, they must stand for retention again.

Prior to each retention election, each justice on the ballot undergoes a review process. This is completed under the auspices of the Colorado Office of Judicial Performance Evaluation. Judicial performance commissions ("Commissions") are required to evaluate each judge based on: surveys of attorneys and non-attorneys, a judge's self-evaluation, interviews with the judge, reviews of opinions/decisions, courtroom observation, and judicial statistics (including caseload information, open cases, case-aging reports, and sentence modifications) (Office of Judicial Performance Evaluation 2012). The Commissions evaluate all judges on the following categories: integrity, legal knowledge, communication skills, judicial temperament, administrative performance, and service to the legal profession/public.

---

[2] See Gill et al. (2011, at 735-36) for a table of the ABA Guidelines and the list of states that use questions measuring each of the categories and subcategories.

The Commissions were created by law in 1988 "for the purpose of providing voters with fair, responsible and constructive evaluations of judges and justices seeking retention" (Office of Judicial Performance Evaluation 2012). There are 22 judicial districts in Colorado, and each district has its own Commission. The individuals sitting on the Commissions are a mixture of both attorneys and non-attorneys and are appointed by the Governor, Chief Justice, Speaker of the House, or the President of the Senate. Each appointed commissioner serves a four year term and can only serve a total of two terms.

When evaluating each judge, the Commissions provide a recommendation of Retain, Do Not Retain, or No Opinion. Each commissioner can cast a vote of either Retain or Do Not Retain. The recommendation of "No Opinion" is only utilized when the Commissions are unable to reach a consensus on a judge or if there is not enough information to make an informed decision. Additionally, if the Commission believes a judge has a significant weakness or area to improve upon, they can recommend a performance improvement plan to address areas of concern.

The information utilized during the evaluation process must be shared with the judge. However, Commission deliberations or interviews are not a matter of public record. Instead, for each judge that is evaluated, the Commission must produce: a short narrative,[3] the final recommendation of the Commission, and the results of the attorney/non-attorney surveys. This is publicly available and can be obtained at the website of the Colorado Office of Judicial Performance Evaluation (http://www.coloradojudicialperformance.gov). Since beginning the

---

[3] The narrative generally includes background information for each judge, such as where they went to law school, the year they were appointed to the court, or any information that may be pertinent to the recommendation. This may include areas of strength/weakness, concerns of the Commission, or whether the judge was placed on a performance improvement plan.

current evaluation process in 1988, the Commissions has evaluated 1,176 judges. They have

assigned "Do Not Retain" recommendations to 17 judges and a total of 10 judges have not been

retained by the voters of Colorado (Office of Judicial Performance Evaluation 2012).

Potential Problems with the Evaluation Process

There are many qualities about the Colorado judicial performance evaluations that

should be commended. It allows for citizen input. It takes into consideration more than just

attorney and non-attorney surveys regarding the quality of judges. Specifically, it allows judges

a chance to evaluate themselves and defend themselves against what could be some biased

attorneys.[4] However, there still remain potential difficulties for the Colorado process.

The development of a comprehensive, state sponsored JPE system was in part a

reaction to the perceived weaknesses of the bar polls that were common in many states at the

time. The independent groups pushing for a more comprehensive JPE system argued that

citizens had little faith in these polls, largely because "many [citizens] especially distrust

lawyers" (Mahoney 1989, 212). But early studies of Colorado's JPE process revealed some

disturbing information about the system. A 1998 American Judicature Society study surveyed

Colorado judges about their impressions of the JPE process (Esterling and Sampson 1998). Only

about 61% of the judges agreed or strongly agreed with the following statements:

1) "I have an adequate opportunity to respond to commission results before they are
   made public."
2) "The overall process used by the evaluation commission to collect information about
   my performance is fair."

Only 30% of judges agreed with a third question:

---

[4] In some instances, the Commission noted that a judge may have scored notably lower in some areas because the survey results received were skewed in favor of either defense or prosecution attorneys.

3) "Judges have access to a fair appeals process if they disagree with the commission's report."

Colorado has attempted to address the concerns expressed by judges in questions 1) and 3) above.[5] Judges are now afforded the opportunity to respond to unfavorable recommendations with a statement of 100 words or less, which the Commission must publish along with the rest of the information. The first such response appears in a 2008 evaluation.

But the fact that these JPE decisions in Colorado happen across a large number of separate Commissions could present consistency problems. As of the current time, there is still not a defined rubric to guide the evaluation process by each Commission. It is important that state JPEs follow "clear rules and procedures for the performance evaluation process" (Andersen 2000, 1388). While all of the Commissions must base their evaluations on a particular list of information sources ("Judicial Performance Fact Sheet" 2012), it is not clear how any one commission distills this information into a rating and corresponding narrative.

The consistency problem is evident quickly when reviewing the narratives produced by the judicial reviews. Some Commissions are very good, providing much background information regarding the judges. This includes law school attended, year the judge was appointed, other major findings from the evaluation process (both positive and negative). Other Commissions include only the barest of facts. Some do not include any information regarding the findings from the evaluation process, whether they are positive or negative. Beyond the "retain" or "do not retain" recommendation, those narratives provide little guidance to a voter. While there would be no justification to conclude the evaluation process is more strenuous or lax in certain

---

[5] These changes are reflected in the "Rules Governing Commissions on Judicial Performance," which can be found here: http://www.coloradojudicialperformance.gov/documents/Rules.pdf.

judicial districts, the lack of consistency in the narratives would indicate there may be

inconsistencies in the evaluation process itself across districts.[6]

As it is currently, the only guidance provided to the Commissions is what information

must be consulted and which qualities to base the judges' evaluation ("Judicial Performance

Fact Sheet" 2012). Each Commission can weigh factors differently when evaluating a judge.

Additionally, the process has a "black box" feel to it. Attorney/non-attorney surveys, judicial

interviews, courtroom visits, etc. are considered by the Commissions and then a

recommendation is presented. It is unclear what information the Commissions are considering,

and which pieces of information weigh more heavily in the final decision. This lack of

transparency in the process again creates concerns over the consistency of the Commissions.

Another area of concern is the very small number of "do not retain" recommendations

is alarming. Only approximately 1% of the total evaluations completed (17 of 1,176) have

resulted in a "do not retain" recommendations. Again, with a more standardized grading rubric,

Commissions would have more guidance when a "do not retain" rating is appropriate. To

illustrate the difficulty of obtaining a recommendation of "do not retain" from the

Commissions, a county judge in 2012 was criticized by approximately 38% of attorneys

surveyed. Specifically cited in most of the surveys was the judge's demeanor. The Commission

was concerned about the judge's use of sarcasm directed toward lawyers who were

unprepared for trial. Furthermore, of the attorneys surveyed who provided an opinion whether

---

[6] There may even be a larger problem with the narratives as it is unclear whether all Commissions are adhering to the "rules" of judicial performance. In a document found on the Colorado Office of Judicial Performance Evaluation's website, effective January 1, 2012, each Commission's narrative was to include the following information: number of commissioners who voted for or against recommendation, undergraduate and law school attended, overall performance of a judge over the evaluation period, and any additional information the Commission feels would help a citizen make a better informed vote choice (http://www.coloradojudicialperformance.gov/ documents/Rules.pdf).

to retain or not retain the judge, the total percentage of attorneys who recommended retention was approximately 20% lower than other county judges. Even still, the committee recommended that this judge be retained.

Another example is a district judge who was recommended for retention by the Commission in 2002. The recommendation was received in spite of the judge's history of losing their temper while on the bench (apparently the 2002 evaluation was not the first time this issue was discussed). However, the Commission believed the judge was committed to improving their temper. Coupled with their satisfaction with the remainder of the judge's performance, the Commission recommended retention in this case as well.

Our purpose for discussing these examples is not to dispute the recommendations provided by the Commissions. However, based on these negative comments, it is unclear what swayed the Commissions toward recommending retention. The examples are provided to serve as an illustration of how non-transparent the evaluation process is. On the flip side of this coin, all but one of the judges who submitted responses to their "do not retain" or "no opinion" evaluation recommendations cited the Commission's selective use of negative information in the final decision. One judge noted that "[t]he Commission disregarded the overall presumptive 'retain' score on my performance survey results." Another judge's frustration is obvious; this judge ends her response with the observation that a "MAJORITY (78%) RECOMMENDS THAT I BE RETAINED."

Judges also express concern about the makeup of the commissions. A judge whose 2002 response was not printed verbatim but was summarized by the Commission noted that the small response rate of the surveys and the absence of prosecuting attorneys from the pool of

attorney respondents may have biased the results. One judge complained that the Commission

was stacked against judges who were tough on crime because it was made up of "the [county]

Public Defender's office chief, the State Public Defender's spouse, a Public Defender's spouse

and a formal Public Defender, a retired Public Defender investigator, a criminal defense

attorney, and no law enforcement representatives."

The materials the Commissions must consider include much of the information

recommended in the ABA Guidelines (American Bar Association 1985, 2005). The Guidelines

seek to help state JPE commissions assemble information that is relevant, quantifiable, and

objective. But the Guidelines are just that: guidelines. There is some evidence that the way

these Guidelines are implemented may fall short of this ideal, especially in terms of the survey

dimension of the process (Gill et al. 2011; Gill 2012). What makes the process of evaluating

Colorado's JPE system more difficult is the fact that there is no set system through which the

various materials available to the Commissions are turned into recommendations.

What is clear is that the Commission recommendation narratives spend a lot of time

talking about the results of the attorney surveys. In the next section, we will evaluate the

quantitative data that the Colorado Office of Judicial Performance Evaluation (OJPE) makes

available on their website, most important of which is the results of the attorney surveys.

THE COLORADO JPE DATA

Although the Commissions do not make public much of the information they use in

making their recommendation decisions, the aggregate results of the survey data is provided to

the public via the OJPE website. These data are available only from 2002 through 2012, even

though the JPE system dates back much farther than that. We have assembled the available

quantitative data from the public reports, and we have supplemented these data with

additional information about the individual judges. We use this information to try to get a

handle on what is driving the JPE results we see in Colorado.

Table 1 presents summary information about the data we have collected. These

summary statistics were calculated on a version of the dataset that is aggregated by judge. For

those judges with more than one evaluation in the dataset, the averages across all of these

evaluations are used. As such, the data in Table 1 are judge-level. Because the Commissions

collect different information about trial judges and appellate judges, we have separated these

groups out for most of our analysis.

Table 1 - Descriptive Statistics Averaged by Judge, Colorado Judges 2002-2012

	All Judges					Appellate Judges					Trial Judges				
	n	mean	Sd	min	max	n	mean	sd	min	max	n	mean	sd	min	max
CJP Retain Recommend	387	0.99	0.10	0	1	29	1	0	1	1	358	0.99	0.09	0	1
Retained by Voters	387	0.99	0.09	0	1	29	1	0	1	1	358	0.99	0.09	0	1
Attorney Retain Score	387	88.26	10.28	41.5	100	29	87.09	5.41	79	96	358	88.35	10.58	41.5	100
Judge Retain Score	-	-	-	-	-	28	97	3.16	87	100	-	-	-	-	-
Layperson Retain Score											358	89.07	8.11	57	100
Attorney Bias Scale	-	-	-	-	-	-	-	-	-	-	321	0.33	0.21	0	1
Female Judge	387	0.28	0.45	0	1	29	0.21	0.41	0	1	358	0.28	0.45	0	1
Minority Judge	387	0.08	0.27	0	1	29	0.10	0.31	0	1	358	0.08	0.26	0	1
Time Since Bar Admit	379	26.80	7.38	5	45	29	30.82	6.77	17	43	350	26.47	7.34	5	45
Law School Ranking	384	2.78	1.09	1	6	29	2.03	0.82	1	3	355	2.85	1.08	1	6
U. Denver Alumnus	383	0.27	0.27	0	1	29	0.17	0.38	0	1	354	0.27	0.45	0	1
Positive or Neutral Story	387	4.13	10.65	0	134	29	6.53	11.51	0	54	358	3.93	10.57	0	134
Negative or Scandal Story	387	0.22	1.11	0	11.5	29	0.47	0.91	0	4	358	0.20	1.12	0	11.5

Retention Information

In Colorado's reported JPE results, the Commissions have given "no opinion" or "do not retain" recommendations only nine times out of the 623 evaluations they have performed; all recipients of these negative recommendations were trial court judges. Only six judges were not retained by the voters, but only two of these were judges who had received negative recommendations from the commissions. This means that only 22% of the judges who got negative Commission recommendations were actually voted out of office, while two-thirds of those voted out of office had "retain" Commission recommendations. Again, all six of the judges voted out of office were trial court judges.

We are also interested in the major survey evaluation components that may influence the committee outcome. All retention scores were obtained from the website for the Colorado Office of Judicial Performance Evaluation. Attorneys were asked "Do you recommend (Judge Name) be retained in office or not retained in office?" The total percentage of attorneys providing a "retain" opinion was utilized.[7] In all, the average attorney retention score is 88.26. Laypeople were asked this same question in their evaluations of trial court judges. Their average score was 89.07. Appellate judges were subject to peer evaluation by other judges; the average peer judge retain score was 97.00.

Additionally, a variable is included that attempts to measure how biased attorneys believe a judge to be. For county and district court judges, attorneys are asked if they believe a

---

[7] For several years there are several sets of retention percentages provided. Through 2002, the survey offered only three options: "Retain," "Do Not Retain," and "No Opinion." Two sets of percentages are computed by the survey report authors: one set includes the no opinion response, the other does not. We use the set that does not include the "No Opinion" category. Beginning in 2004, the survey began to offer a range of responses: "strongly not retain," "somewhat not retain," "undecided," "somewhat retain," and "strongly retain." For the surveys from 2004-2012, we've used the sum of "somewhat retain" and "strongly retain" to calculate the retention score.

judge is biased toward the prosecution or defense. To create a measure of bias, the aggregate

percentage of attorneys who believed the judge to be biased was offset against each other to

obtain a "net bias" score. This score was folded on itself such that higher levels of bias toward

either the prosecution or the defense are closer to the maximum value of 1, and less bias is

closer to the minimum value of zero. The trial court judges overall have an average score of .33

out of one. Unfortunately, a similar score could not be created for the appellate court judges,

as this question was not asked of respondents evaluating Court of Appeals or Supreme Court

judges.

We have also collected a number of other pieces of information about the judges in the

hopes of explaining the outcome of the JPEs. Of particular interest in this research is

determining whether any implicit gender bias is skewing the results of the evaluation process.

As such, gender and minority status were coded based on pictures included in the

biographies/narratives of each judge. While there are a substantial number of female judges in

Colorado (28% of 358 judges), there are very few minority judges in the analysis (7.54% of 358

judges). We have also included the number of years elapsed since the judge was admitted to

the bar.[8] The average number of years since bar admittance is 26.8 years, although there is a

wide range of values (5-45 years).

A measure of the prestige of the judge's law school is also included in the models. Law

school alma maters were determined either through the judicial narratives provided by the

judicial performance commissions or through Martindale. Based on the 2012 US News and

---

[8] In those instances where the judge is a lay judge that was not admitted to the bar, we use the year in which the judge completed his or her graduate education, where appropriate. When the lay judge had no graduate degree, we set the date at three years past the attainment of the bachelor's degree.

World Report Law School Rankings, all law schools were assigned a category. Those schools ranked 1-15 by US News and World Report were coded as a 1, 16-50 – 2, 51-100 – 3, 101-150 – 4, 151+/unranked – 5. If an individual did not attend a law school, they were assigned a score of 6. The distribution of law school prestige is presented in Figure 1. We have also created a series of dummy variables for judges who graduated from the University of Denver Sturm College of Law (27.32% of judges) and the University of Colorado Law School (26.20% of judges), the only two law schools in Colorado. In all, slightly more than half of the judges received JDs from Colorado law schools.

Figure 1 - U.S. News Rankings of Law Schools Attended by Colorado Judges, 2002-2012
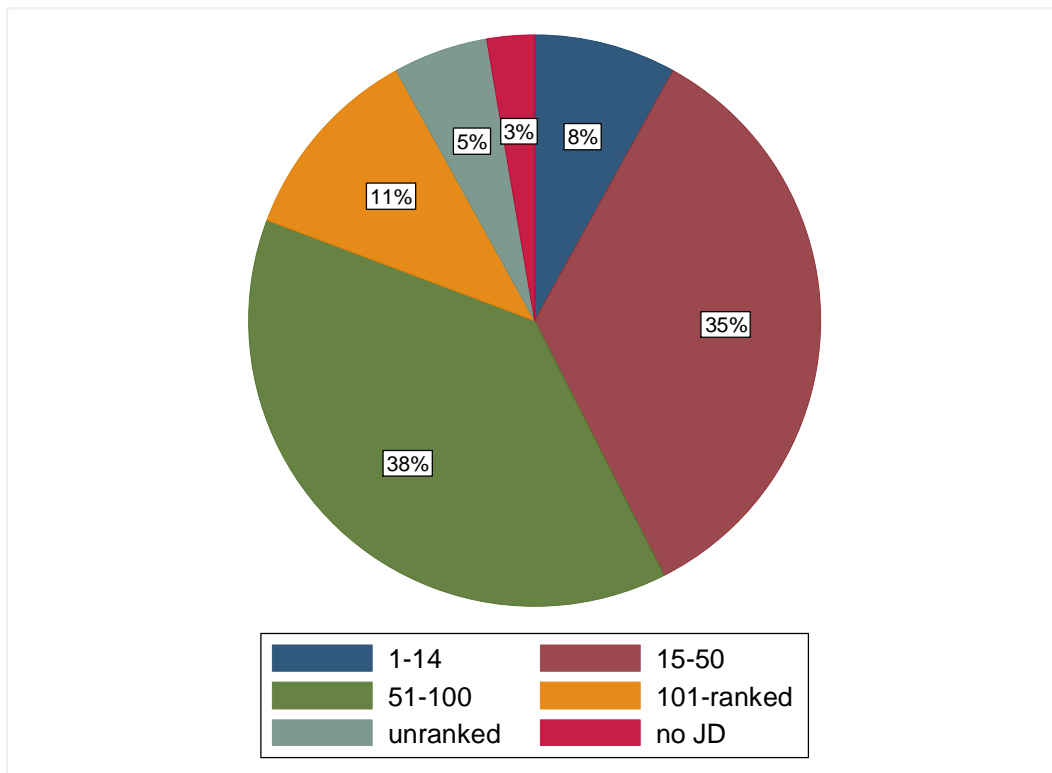
Because there may be a relationship between the number of times a judge is mentioned in the news and the chances of retention, a variable is included that indicates how many times each judge was mentioned in the Denver Post. For each evaluation period, only the time from

the previous evaluation to 45 days prior to the current election are included. This is because the

judicial performance evaluations must be completed prior to 45 days before the election. Any

press received after the evaluations are completed and released to the public would/could not

influence a lawyer's evaluation because that evaluation survey would have already been

submitted.

We assume that how a judge is portrayed in a news story could impact a lawyer's

perception of that judge. Because of this, news stories were classified as neutral/positive or

negative. Those that were positive either praised a ruling or discussed how good of a job a

judge is doing. Neutral press mentions the judge, but takes no stance for or against them. In our

sample, judges had an average of 4 positive or neutral mentions in the press, with a maximum

of 134 stories. Negative stories criticized the judge's decisions, connected the judge to a scandal

of some sort, or were otherwise openly hostile to the judge. There were fewer of these kinds of

stories; the average number of negative stories was .22, although the number was twice that

for appellate judges.

MODELING JPE RESULTS

Our goal in this preliminary analysis of Colorado JPE data is to get a sense of what the

JPE process is measuring. The first step is to get a sense of what pieces of information influence

the probability that a judge will get a positive recommendation from the JPE Commission. To

investigate this, we performed an exploratory analysis using various explanatory variables to

predict the positive recommendation. Our dataset is an unbalanced panel, since we have

multiple evaluations for many of the judges, but not for all judges. We are particularly

interested in finding out how heavily the commissions rely on the information from the various

survey components of the JPEs. Because only the trial judges were evaluated by both attorneys

and laypeople, a model comparing the effects of these two different sources of information can

only include the universe of trial court judges.

Table 2 - Population Averaged Generalized Estimating Equations Model of Positive JPE Recommendations

	Odds Ratio	SE	z	P > |z|
Female Judge	0.205	0.166	-1.96	.051
Years Since Bar Admission	0.975	0.056	-0.43	.664
University of Denver JD	1.456	1.254	0.44	.662
Law School Prestige	0.758	0.253	-0.83	.406
Neutral/Positive Stories	1.090	0.145	0.65	.518
Critical/Scandal Stories	0.900	0.103	-0.93	.354
Attorney Retention Score	1.076	0.025	3.20	.001
Layperson Retention Score	1.041	0.039	1.05	.292
Constant	0.036	0.132	-0.91	.362
N=575 evaluations of 349 judges; evaluations per judge mean = 1.6, min=1, max=5; Wald χ^2=18.85, p=.015				

The results of this model can be found in Table 2. The covariates in this model do not

help to predict the probability of a positive Commission recommendation, with two exceptions.

Although the layperson retention scores do not help to predict positive recommendations, the

attorney retention scores do. This suggests that the Commissions may be relying on the surveys

of attorneys to help them establish a baseline for their decision to give a negative performance

evaluation.

The other significant predictor of positive recommendations in our model is judge sex.

Female judges have a lower likelihood, all other things equal, of receiving a negative

recommendation from the Committees. The plot in Figure 2 illustrates the effect that having a

low attorney survey score has for men and women. The result suggests that the Committees

might be slightly less willing to give women judges with poor attorney retention scores a

positive recommendation when compared to their similarly situated male colleagues. **Figure 2**

22

Table 3 - Predicted Marginal Probabilities of Positive Committee Recommendation by Sex over Attorney Score
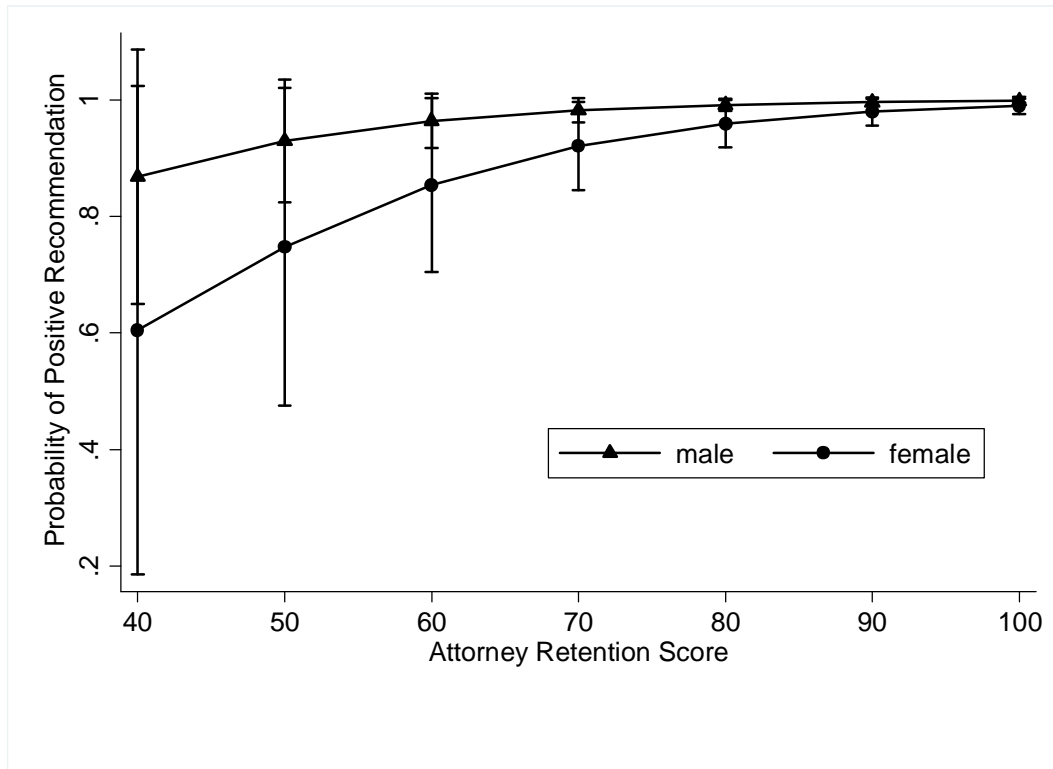
Since the attorney survey is such an important predictor of negative Commission recommendations, we investigate the makeup of these scores. To estimate the effects of judge characteristics on the resulting attorney survey scores, we use a pooled ordinary least squares regression with panel corrected standard errors (Beck and Katz 1995). This method allows us to control for the fact that the repeated observations of each judge are not independent from one another. It is also a good way to deal with an unbalanced panel, which we have because not all judges were evaluated in each evaluation cycle in our data.

The Colorado JPEs do not ask the same questions of respondents evaluating appellate judges as they do of respondents evaluating trial judges. For this reason, we analyze the results of the appellate judges separately from those of the trial judges. We find one main similarity

common to the models, along with a couple of interesting differences. The results of the model

of attorney retention scores for appellate judges can be found in Table 4.

Table 4 Pooled OLS Regression with Panel Corrected Standard Errors of Attorney Retention Scores for Appellate Judges

	Coef.	PCSE	z	P > |z|
Female Judge	-3.042	1.524	-2.00	.046
Years Since Bar Admission	-0.288	0.131	-2.20	.028
University of Denver JD	6.249	1.693	3.69	.000
Law School Prestige	-1.736	1.405	-1.24	.216
Neutral/Positive Stories	0.364	0.114	3.18	.001
Critical/Scandal Stories	-3.693	1.418	-2.60	.009
Constant	97.963	5.231	18.73	.000
N=48 evaluations of 29 judges; evaluations per judge mean = 1.655, min=1, max=4; r^2=.30; Wald χ^2=63.88, p=.000				

Our model shows that, all else equal, female judges score significantly lower than male

judges. In our model, the only variable that is not significant is the prestige of the judge's law

school. However, appellate judges get a large boost when they have graduated with their JD

from the University of Denver Sturm College of Law. Interestingly, alternative specifications for

local connections, including a measure of having attended the University of Colorado School of

Law and a measure of having attended either of these Colorado law schools, were not

significant. As we would expect, the attorney evaluations are influenced by the media coverage

of the judges. Positive or neutral stories boost scores by about a third of a point per story.

Critical stories or stories linking the judge to a scandal are quite harmful; these reduce the

judge's score by 3.7 points per story.

Table 5 - Pooled OLS Regression with Panel Corrected Standard Errors of Attorney Retention Scores for Trial Judges

	Coef.	PCSE	z	P > |z|
Female Judge	-3.193	0.493	-6.47	.000
Years Since Bar Admission	0.041	0.043	0.94	.346
University of Denver JD	-0.435	0.893	-0.49	.626
Law School Prestige	-0.192	0.418	-0.46	.645
Neutral/Positive Stories	-0.095	0.020	-4.66	.000
Critical/Scandal Stories	0.024	0.317	0.08	.939
Bias Score	-14.947	2.143	-6.98	.000
Constant	94.510	2.736	34.54	.000
N=614 evaluations of 314 judges; evals per judge mean = 1.955, min=1, max=6; r^2=.13; Wald χ^2=217.63, p=.000				

The results of our analysis of trial judges appear in Table 5. As with the appellate judges, female judges score significantly lower than their male counterparts. Most of the other measures that we hypothesized would be driving the attorney retention scores of trial court judges are not significant. Aside from judge gender, only two of the other measures have significant explanatory power. The first of these is the category of positive or neutral coverage in the news. Interestingly, this media coverage actually decreases the predicted scores for the judges. The other significant independent variable is the attorneys' assessment of the degree to which the judge is biased for the prosecution or the defense. Moving from the lowest value to the highest value of this variable decreases the retention score by 15 points. This relationship, along with the sex-based difference in predicted scores, is presented graphically in Figure 2.

DISCUSSION

The results of this analysis show a statistically significant gender penalty for female judges of about 3 points in the attorney surveys. This is a small difference, it is true. But we believe that this pattern is something that still deserves attention. It is likely that this small but persistent difference in scores reflects a difference in the way that female judges experience

Figure 2 - Predicted Marginal Effect of Judge Sex over Values of Bias Scores for Trial Court Judges
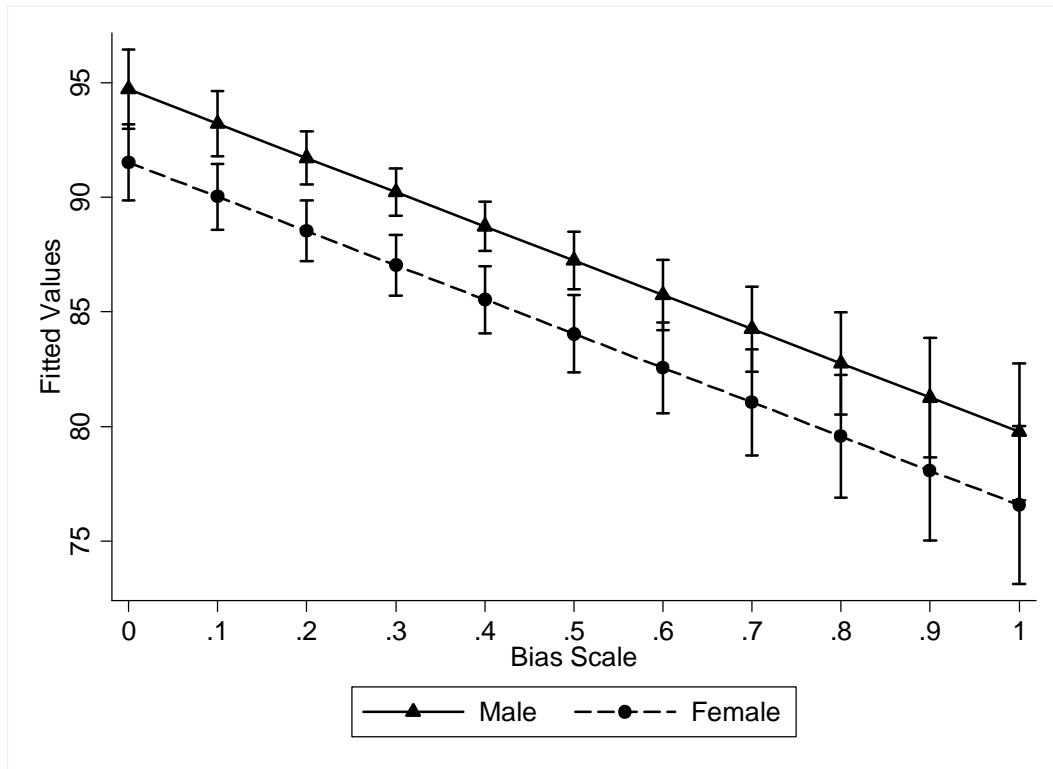
their work on a day-to-day basis. That such a bias is reflected in the attorney surveys suggests that the bias may be reflected in the way that female judges are treated in their courtrooms (Rhode 2001; Bazelon 2009).

Another reason to take this disparity seriously is the fact that the Commission decision relies heavily on these attorney surveys. Worse still, the Commissions are also significantly more likely to give female judges a negative recommendation even after controlling for the implicit bias already present in the attorney surveys. While the impact of this bias is of a relatively small magnitude at the attorney survey level, it appears that this gender gap is multiplied at the Commission level, where commissioners add their own implicit biases on top of those already contained in the attorney surveys.

The obvious question now is what to do about this problem. Previous research on gender bias in attorney surveys conducted by non-state actors (Gill 2012; Gill et al. 2011; Wood and Lazos 2010) was largely dismissed by proponents of merit selection and JPE (Kourlis 2010) on the basis that state-sponsored programs are unlikely to suffer the same methodological as these "bar polls" (Andersen 2000). It is true that Colorado's program is much more comprehensive than just a survey of attorneys. Colorado has taken pains to get its Commissioners to consider information from a large variety of sources in its decision making process. If done systematically and with careful planning to avoid the problem of implicit bias, this can be a very effective strategy.

But the problem of implicit bias persists, even in Colorado's comprehensive and well-funded JPE system. It is difficult to identify confidently a single source of this bias; it is likely coming from a number of the evaluation criteria the Commissions are charged with considering when making their recommendations. A large part of the problem, however, is that it is difficult to know the source of the bias. The Commissions do not follow a set procedure for weighing the evidence, and this preliminary analysis suggests that the attorney surveys play a substantial role in at least framing the Commissions' recommendations. And, although the attorney surveys are conducted in connection with a state-sponsored JPE program, they are still virtually indistinguishable from the unofficial "bar poll" surveys in other states (Gill 2012).

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.

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

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

Bazelon, Emily. 2009. "The Bitch Defense." In *XXfactor: What Women Really Think*: Slate.

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.

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

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

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.

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*, ed. L. Nadel. 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.

Esterling, Kevin M., and Kathleen M. Sampson. 1998. *Judicial Retention Evaluation Programs in Four States: A Report with Recommendations*. Chicago, IL: American Judicature Society.

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):1049-160.

Gill, Rebecca D. 2012. "Judicial Performance Evaluations as Biased and Invalid Measures: Why the ABA Guidelines are Not Good Enough." In *Midwest Political Science Association*. Chicago, IL.

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 Blac Lawyer Sit at Our Table." *Fordham Law Review* 71 (4):1257.

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.

"Judicial Performance Fact Sheet." 2012. ed. O. o. J. P. Evaluation. Denver, CO.

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." In *Quality Judges Initiative*. Denver, CO: Institute for the Advancement of the American Legal System.

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

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.

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

Mahoney, Anne Rankin. 1989. "Citizen Evaluation of Judicial Performance: The Colorado Experience." *Judicature* 72 (4):210-6.

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

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

White, Penny J. 2001. "Judging Judges: Securing Judicial Independence by Use of Judicial Performance Evaluations." *Fordham Law Review* 29:1053.

———. 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." In *William S. Boyd School of Law Research Papers*. Las Vegas, NV.

———. 2010. "Gender and Race Bias in the Proposed Judicial Performance Evaluations: Some Preliminary Findings." In *William S. Boyd School of Law Research Papers*. Las Vegas, NV.