Bias in Open Peer-Review: Evidence from the English Superior Courts[*]

Jordi Blanes i Vidal[†] Clare Leaver[‡]

July 10, 2014

Abstract

This paper explores possible biases in open peer-review using data from the English superior courts. Exploiting the random timing of on-the-job interaction between reviewers and reviewees, we find evidence that reviewers are reluctant to reverse the judgements of reviewees with whom they are about to interact, and that this effect is stronger when reviewer and reviewee share the same rank. The average bias is substantial: the proportion of reviewer affirmances is 30 percentage points higher in the group where reviewers know they will soon work with their reviewee, relative to groups where such interaction is absent. Our results suggest reforms for the judicial listing process, and caution against recent trends in performance appraisal techniques and scientific publishing.

Keywords: courts and judges, open peer-review, workplace interactions.
JEL Classification: A12, C21, K40, Z13.

[*] An early version of this paper was circulated under the title 'Using Group Transitions to Estimate the Effect of Social Interactions on Judicial Decisions'. We are grateful to audiences at Oxford, Bristol, LSE, Munich, Frankfurt, Ammersee (Workshop on Field and Natural Experiments), Copenhagen (Microeconometrics Network), UPF, Amsterdam Center for Law and Economics, Aberdeen, and Leicester for insightful comments and suggestions, and HM Courts and Tribunals Service and the Ministry of Justice for many helpful conversations. Specific thanks are due to Ian Crawford, Ian Jewitt, Ignacio Palacios-Huerta, Steve Pischke, Rocco Macchiavello, James Rockey, Yona Rubinstein, Sarah Smith, Daniel Sturm, Chris Wallace, Michael Whinston, the Editor and three referees. All errors are our own.

[†] London School of Economics. Email: j.blanes-i-vidal@lse.ac.uk.

[‡] University of Oxford, CEPR and CMPO. Email: clare.leaver@economics.ox.ac.uk.

1 Introduction

Peer-review is defined as the evaluation of a person's work by a group of people in the same field. Although the term is typically associated with scientific publishing, similar practices are used in many professional settings. When evaluating applications for funding, the U.S. National Science Foundation and the U.K. Economic and Social Research Council solicit reviews from researchers working in the applicants' fields of study. Away from academia, the legislative branch of the U.K. government solicits reports on the work of ministers (typically MPs) and their departments from a select committee of other MPs. When civil appeals are granted, the judicial branch of government solicits opinions on the work of first instance judges from panels of other judges sitting in appellate courts. In professional service firms, performance appraisal systems are often based on evaluations by fellow employees.

One reason why peer-review is so pervasive in such settings is that fellow professionals are thought to be better placed to offer informed assessments than non-experts. While there is some evidence to support this view (Kassirer and Campion 1994), there are also potential disadvantages to using peer-review. One potential downside arises precisely because peers are experts. Since the rationale for using an expert is that the assessment will proceed subjectively (from the expert's mind), the outcome of the review process could be affected by chance (Cole, Cole and Simon 1981) and/or discrimination on the basis of personal characteristics (Peters and Ceci 1982, Gilbert, Williams and Lundberg 1994, Ginther et al 2011). A second potential downside arises because peers working in the same narrow field may have met while training at the same institution (Blanes i Vidal and Leaver 2011) or collaborating while on the job (Fafchamps, Goyal and van der Leij 2010, Blanes i Vidal and Leaver 2013). When an evaluation is undertaken by a reviewer with a personal tie to the reviewee, the outcome could be subject to favouritism based on friendship (Wenneras and Wold 1997) or familiarity (Li 2012).

It has been argued that these potential downsides are specific to the traditional single-blind system, where reviewers remain anonymous but are aware of the identity of their reviewees. In particular, supporters of an alternative double-blind system claim that anonymising the identity of the reviewees minimises the chances of reviewer bias (Blank 1991). In fact, the evidence from numerous randomised controlled trials of double versus single-blind reviewing does not unequivocally support the view that blinding the identity of reviewees improves the quality of reviews (Smith 1999). Moreover, in many contexts blind reviewing is either impractical or indefensible on ethical grounds.[1]

[1] For instance, in scientific publishing internet searches can quickly remove author anonymity, while ethical considerations ensure that judicial hearings are open in most democracies ("a court with an unidentified judge makes us think immediately of totalitarian states and the world of Franz Kafka", Smith 1999 p. 4).

For these reasons, an open system, where the identities of the reviewers and reviewees are public, has received attention. Proponents of open peer-review claim that removing the anonymity of reviewers has ethical and intellectual benefits and, by fostering reputational accountability, also minimises reviewer bias (Robertson 1976, Fabiato 1994, Goldlee 2002).[2] Most proponents acknowledge that open reviewing could, in principle, lead to alternative forms of bias −e.g. reviewers, feeling obliged to justify negative comments, might "take the easy way out" by issuing a positive review, while reviewers in workplace networks might favour "people in their group expecting reciprocity"(Fabiato 1994)− but typically dismiss these possibilities on *a priori* grounds. Robertson (1976), for instance, acknowledges that under an open system reviewers might "fear making enemies among friends and influential colleagues, and that this would lead to a 'kid gloves' approach" but he dismisses this possibility because it means "taking the somewhat cynical and paternalistic view that a scientist's commitment to objective truth would give way far too often to his prejudices and ambitions".

The efficacy of open peer-review is an empirical question, however. If reputational accountability is strong, reviews may be unbiased; but, if it is weak, there could be discrimination, backward-looking favouritism motivated by pre-existing personal ties, forward-looking favouritism driven by a fear of future awkwardness and/or reprisal, or all three. There have been few attempts to investigate this issue empirically. To the best of our knowledge, only a small number of randomised control trials of open versus blind reviewing have been conducted to date and, typically, these studies have not been designed to elicit the mechanism behind any potential effect. The objective of this paper is to explore whether open peer-review *is* subject to bias and, if so, to highlight the underlying economic mechanism.

We study panels of judges, sitting in the English Court of Appeal, reviewing judgements taken by other judges sitting in the High Court. We choose this setting because testing for bias in the English superior courts is a worthwhile exercise in its own right, and because institutional features of these courts can be used to isolate the mechanism behind any potential effect. Focusing on a judicial setting does have a disadvantage, however: the doctrine of natural justice prohibits blinding of reviewers and reviewees. Although this means that we cannot directly compare open versus blind reviewing, policy-relevant lessons can still be learnt from our analysis. For instance, evidence of backward-looking favouritism bias would suggest a policy aimed at weakening existing ties between reviewers and reviewees (e.g. via a conflict of interest test at the time of the review), while evidence of forward-looking favouritism bias would suggest a policy aimed at limiting *future* links between reviewers and reviewees (e.g. by increasing the distance between them in the judicial hierarchy). Moreover, such insights would be valuable in other

[2] In an early contribution to the debate, Robertson (1976, p. 410) suggests that "if a referee's identity is known, his professional reputation is directly at stake and so he would take more time and care before passing judgement".

professional settings, such as performance appraisals and scientific publishing, where the usage of open peer-review is growing.[3]

Our empirical strategy exploits variation in on-the-job interaction between reviewers and reviewees. An observation is a 'panel-reviewed judge' pair. Each panel reviews a judgement taken by a judge sitting in the High Court and must decide whether to affirm this judgement, or to reverse it indicating that the reviewed judge was wrong on a point of law. A reversal has detrimental consequences for the reviewed judge, e.g. by reducing his chances of promotion (The Judges' Council 2003).[4] On-the-job interaction occurs when a panel member works together with the reviewed judge on an *unrelated* appeal.

This setting enables us to test for two of the sources of bias noted above, backward and foward-looking favouritism.[5] The sociology literature suggests that interaction occurring in a situation of cooperative interdependence, such as working together on an appeal, is likely to promote friendship (Moody 2001). If backward-looking favouritism exists, we should therefore see a higher affirmance rate when a panel member has worked with the reviewed judge than when all panel members lack on-the-job interaction. Equally, working together on an appeal is an opportunity to confront a panel member for a past reversal and to seek revenge via uncollegial behaviour (Cross and Tiller 2008). Fears of awkwardness and reprisal are likely to loom large prior to a meeting. If forward-looking favouritism is an important force, we should therefore see a higher affirmance rate when a panel member knows he is about to work with the reviewed judge (who will be aware of the review decision) than when all panel members know such interaction is not about to occur.

To use these observations, we require an exogenous source of variation in on-the-job interaction. Unfortunately, as we explain in Section 2, the *level* of on-the-job interaction between reviewers and reviewees is likely to be correlated with unobserved selection variables. This is because a panel member can only experience on-the-job interaction if the senior judiciary deems the reviewed judge to be of

[3] As Murphy and Cleveland (1995) note, the latter half of last century saw two trends in performance appraisal techniques: reviews were more likely to be open (available to the employee) and decentralised (conducted by the employee's immediate line manager rather than upper-level management). More recently, '360 degree' reviews based on assessments by customers, subordinates and peers, as well as managers, have become popular. In scientific publishing, the BioMed Central journals and the *British Medical Journal* pioneered the use of open peer-review in 1999 and continue to use it today. *Nature* and *PLoS Medicine* experimented with a voluntary system (where reviewers could choose to sign reports) in the mid-2000s but discontinued this practice due to low take-up. Since then, open peer-review has been adopted in the physical sciences, including the leading journal *Atmospheric Chemistry and Physics* and other open-access journals published by the European Geosciences Union. In 2012, a leading humanities journal, *Shakespeare Quarterly*, put together a special issue using open peer-review. In 2013, a new journal in the biological and medical sciences, *PeerJ*, adopted open peer-review alongside an innovative 'fixed fee' business model.

[4] Both Salzberger and Fenn (1999) and Blanes i Vidal and Leaver (2011) document that reversals are negatively associated with promotion prospects in the English superior courts.

[5] Data limitations prevent us from testing for discrimination.

sufficient ability to sit on the appellate bench, and such unobserved perceptions of ability will almost certainly correlate with the panel's decision to affirm or reverse the reviewed judge's first instance judgement.

We respond to this selection problem by employing a methodology that utilises variation in the *order* of a given level of treatment (i.e. whether a panel member works with the reviewed judge at date t but not at date s, or vice versa). In Section 3, we show that, under two plausible symmetry assumptions on the joint distribution of the treatment and selection variables (and for sufficiently small $s-t$), we can identify the average effect of treatment order for units treated once. The logic behind this identification strategy is simple: under our symmetry assumptions, the order of treatment is random conditional on unobservables staying fixed over time, and this is almost certainly the case when $s-t$ is sufficiently small. In other words, perceptions of ability may well determine whether the reviewed judge *ever* sits on the appellate bench but not whether this happens today rather than tomorrow.

This insight enables us to proceed to an estimation via a comparison of means. Specifically, we compute the difference between the mean affirmance rate for panels aware of an interaction before, but not after, the review decision and the mean affirmance rate for panels aware of an interaction after, but not before, the review decision. We argue that, if this difference is positive (respectively negative), we can reject the hypothesis that panels are above the influence of on-the-job interaction *and* conclude that the predominant force is backward-looking favouritism motivated by personal ties (respectively forward-looking favouritism driven by a fear of awkwardness and/or reprisal).

In Section 4, we explain how our comparison groups are constructed using 10-day periods before and after the review, as well as the regression models that we use to perform robustness checks. Our main results are presented in Section 5. The key finding is that the mean affirmance rate for panels with an interaction in the 10 days before the review but not in the 10 days after the review is significantly *smaller* (by 30 percentage points) than the mean affirmance rate for panels with an interaction in the 10 days after the review but not in the 10 days before the review. The magnitude of this effect is robust to controlling for an array of observable characteristics, as well as for treatment in other periods.

We interpret the finding that anticipated interaction increases the affirmance rate as evidence that, when lacking the protection of anonymity, reviewers may indeed take a lenient "kid gloves" approach. In Section 6, we assess this mechanism in greater detail. First, we substantiate the rationale for forward-looking favouritism by providing evidence that uncollegial behaviour is lower during on-the-job interactions that occur after, rather than before, an affirmance (but not after, rather than before, a reversal). Next, we show that reviewers suffer less from forward-looking favouritism bias when assessing junior colleagues than when assessing peers of the same rank. Finally, we draw out additional empirical predictions relating to the *quality* of review decisions. Developing a simple theoretical framework, we show how an anticipated on-the-job interaction can cause a forward-looking favouritism bias that: (a) increases the probability that the review decision is incorrect; (b) increases the probability that an affirmance is incorrect; and (c) decreases the probability that a reversal is incorrect. Using data on legal challenges of review decisions to the House of Lords, and citations of review decisions by other judges, we find evidence to support these predictions. A legal challenge to the House of Lords is significantly *less* likely among panels that reverse their reviewed judge in advance of an anticipated on-the-job interaction than among panels that reverse prior to an unanticipated on-the-job interaction. Moreover, the difference in the effect of an anticipated interaction on the likelihood of a legal challenge when the review decision is an affirmance rather than a reversal is positive and strongly significant.

In light of these results, we argue that, contrary to previous claims (e.g. Robertson 1976), open peer-review can be subject to favouritism bias. The obvious policy lesson is for the English superior courts. HM Courts and Tribunals Service should consider reforming the listing process to ensure that judges cannot anticipate that they will soon sit with colleagues affected by their decisions. As we explain in Section 7, this could be achieved by limiting the downward movement of judges (to increase the distance between reviewers and reviewees in the judicial hierarchy) or, more laboriously, by vetting potential panels for the presence of a reviewer-reviewee pair. There are also lessons for other settings. Our finding that reviewers suffer less from forward-looking favouritism bias when assessing junior colleagues suggests that firms should reconsider the merits of *decentralised* open performance appraisals, and highlights the need for anonymity in multi-rater '360 degree' reviews. Our results also provide econometric support for a submission to the U.K. Government's recent investigation into peer-review in scientific publications, namely that open peer-review may only be suitable in broad fields where reviewers and reviewees "don't bump into each other the next day" (Science and Technology Committee 2011, Paragraph 19).

Related Literature A small number of studies have investigated the efficacy of open peer-review by randomly assigning journal reviewers to an open or single-blind treatment. Echoing our result, Walsh et al (2000) report that reviewers for the *British Journal of Psychiatry* were more likely to recommend acceptance under the open treatment. Godlee et al (1998) and van Rooyen et al (1999) report that reviewers for the *British Medical Journal* showed no differences in acceptance rates across treatments but, in the latter study, invitations to review were more likely to be declined. There have also been attempts to investigate this issue within performance appraisal systems. Antonioni (1994) reports that reviewers rated their reviewee more highly under an open treatment where the appraisal questionnaire required the reviewer to identify him/herself, than under a single-blind treatment where the appraisal questionnaire was anonymous. Similar studies have found evidence of 'rating inflation' within open performance appraisals in the teaching and nursing professions (Afonso et al 2005, Kagan et al 2006).

Turning to our judicial application, there have been numerous studies of decision-making in the U.S. Courts of Appeals. These studies explore whether the political ideologies or backgrounds of appellate judges influence case outcomes (see Sisk, Heise and Morriss 1998, Sunstein et al 2006 and the references therein). To the best of our knowledge, no study has examined whether appellate panels are swayed by the characteristics of (or personal contact with) the federal district judges that they are reviewing.[6] There has been little empirical work on decision-making in our setting, the English Court of Appeal. Two exceptions are Blackwell (2011), who looks for panel effects in immigration and employment cases, and Blanes i Vidal and Leaver (2013), who explore whether appellate panels are influenced by a strategic desire to cite previous appeal judgements. Again, neither paper investigates whether appellate panels are swayed by the characteristics of the judges that they are reviewing.

From a more methodological perspective, our paper contributes to the literature on treatment evaluation (see, e.g., Imbens and Wooldridge 2009). The empirical strategy set out in Section 3 bears some similarity to both symmetric differences-in-differences and regression discontinuity design (Lee and Lemieux 2010). Symmetric differences-in-differences estimation exploits the fact that, for each unit of analysis, the outcome of interest is observed at two dates (before, and an equidistant time after, a single selection decision). The key statistical assumption is that any unobserved transitory component of the outcome is covariance stationary. Regression discontinuity design, on the other hand, exploits the fact that a 'threshold' selection variable is observed for each unit of analysis, and sufficiently many units fall arbitrarily close to this threshold. The key statistical assumption is that the conditional mean of any *un*observed selection variable is continuous at the threshold. In contrast to these approaches, our research design exploits the fact that, for each unit of analysis, treatment status is observed at two dates arbitrarily close in time (which could, but need not, be just before and after the single outcome of interest). The key statistical assumptions are that the propensity score function is stationary and unobserved selection variables follow a Markov process with a symmetric transition rate matrix.

[6] Steinbuch (2009) looks for, and finds, a correlation between the political ideology of district court judges and the likelihood of reversal by the U.S. Court of Appeals for the Eighth Circuit. However, as he admits, his empirical strategy cannot ascertain whether this correlation is caused by a disparity in the world view of judges at different tiers of the judicial hierarchy, or bias against district court judges belonging to a particular political party. Choi, Gulati and Posner (2010) pose the reverse question and explore whether district court judges are swayed by appellate decision-making. Epstein et al. (2009) document that U.S. Supreme Court judges tend to disproportionately affirm cases appealed from the circuit where they have previously served.

2 Institutional Background

Our study is based on reviews of judgements in the English superior courts. These judgements are taken by judges sitting (alone) in the High Court, while the reviews are undertaken by judges sitting in panels (of two or three) in the Civil Division of the Court of Appeal (hereafter the CA Civ). In this Section, we explain how these two groups of judges may come to work together on an unrelated appeal.[7]

The panels are formed by a bureaucrat known as a Listing Officer. Once a litigant has been given leave to appeal, the Listing Officer establishes how many panel members are required and then applies the rule that each panel member should be drawn from the list of *ticketed* judges (those allowed to sit in the CA Civ) in accordance with the *cab-rank principle*. We will say that a review panel is treated at a given date if a reviewer from this panel experiences an on-the-job interaction with the reviewed judge on this date. So defined, the probability of treatment at a given date depends on three processes. The first determines the number of reviewers, the second whether the reviewed judge and any of these reviewers are ticketed at the given date, and the third whether, conditional on being ticketed, the reviewed judge and a reviewer are actually matched at this date.

The number of reviewers is governed by statute. Some legal subjects can be reviewed by two judges, but most will require three judges. Since some legal subjects (e.g. public and administrative law) are known to be prone to reversals, the number of reviewers is a candidate selection variable, correlated with both the likelihood that a panel is treated with an interaction and its propensity to reverse.[8]

The list of ticketed judges is chosen by the senior judiciary. Judges serving in the post of Lord Justice or Law Lord are automatically ticketed. Promotions in the English Senior Judiciary are determined by perceived quality, experience and legal specialism (Blanes i Vidal and Leaver 2011). In contrast to the U.S., political affiliations seem to play at most a minor role (Griffith 1997, Robertson 1998). Judges serving in the more junior post of Justice (and retired Justices, retired Lord Justices, and retired Law Lords) can be ticketed but this is at the discretion of the Head of Civil Justice. As Table 1 illustrates, a similar ticketing rule applies to the High Court. These rules suggest that a number of factors are likely to influence the ticketing process. For reviewed judges who held the post of Justice at the time of their judgement, an important factor will be the senior judiciary's perception of their quality.[9] In contrast, reviewed judges who held the post of Lord Justice at the time of their judgement will be automatically

[7] Readers unfamiliar with the English system may find it helpful to refer to Table 1 prior to reading this Section. See also Blanes i Vidal and Leaver (2011) for a more detailed summary of the institutional details of these courts, as well as a discussion of other explicitly social forms of interaction within the senior English judiciary.

[8] Note that the legal subject is determined at the first instance stage.

[9] To be ticketed, these reviewed judges need to have impressed either the Head of Civil Justice or the committee in charge of promotions to the post of Lord Justice.

ticketed unless they have retired, and so an important factor will be their age. Since the rank, perceived quality and age of the reviewed judge are likely to influence the panel's decision, ticketing status is also a candidate selection variable. Unfortunately, historical lists of ticketed judges are unavailable, and so we do not observe this candidate selection variable (or correlates such as perceived quality) for all of the relevant judges.

The cab-rank principle works just as its name suggests: judges completing a review join the back of the queue; when a new review requiring a panel of size n arrives, the bureaucrat allocates it to the n judges closest to the front of the queue. At the start of a legal term (or within a term where reviews have been completed at the same time) there will be more than n judges in the first position of the queue. In the event of such a tie, the panel is formed at random.[10] During the rest of the term, judges join and leave the cab-rank at a high frequency. This is because, with the discussion limited to points of law, reviews are completed quickly, typically in just a few days.[11] Our empirical strategy exploits the fact that matches between *ticketed* judges are random (by virtue of the cab-rank principle) and highly frequent (by virtue of being appeals) to solve the problem of selection on unobservables.[12]

To test the forward-looking favouritism hypothesis, we make use of two further features of the CA Civ, namely that during our sample period panels were typically listed one month in advance of the hearing,[13] and hearings were open and immediately summarised in newspaper law reports.[14] Thus, when taking their review decision, panel members should know for certain whether they will or will not work with the reviewed judge within the next 30 days *and* anticipate that, during any such interaction, the reviewed judge will be aware of their decision.

[10] See Blanes i Vidal and Leaver (2013) for a formal test (and confirmation) of this claim.

[11] An appeal heard in the CA Civ may concern questions of fact, award of damages, exercises of discretion, questions of law, or a request for a new trial in the lower court. The standard of deference given to the High Court is high, implying that the reviews in our dataset will typically only concern exercises of discretion or questions of law. Even in the former case, the CA Civ will only interfere with the judge's discretion if he/she has "erred in law, applied an incorrect principle, misapprehended the facts, taken irrelevant matters into consideration or ignored relevant considerations, or if the court is satisfied that the decision is wrong" Bailey el al (2002, p. 1300).

[12] Note that 'quicker' judges will join the back of the queue more often and will therefore accumulate more interactions. This fact underlines the need to use the empirical strategy set out in Section 3.

[13] Information on the timing of current CA Civ listings can be obtained from HM Courts and Tribunals Service. According to a Listing officer that we spoke to: CA Civ listings are updated on a daily basis; typically, the composition of the panel is public information at least one month in advance; and, while changes in the composition of the panel can occur, they are rare and unlikely to happen shortly before a review. Unfortunately, there are no historical records documenting exactly how far in advance of each of the reviews in our sample it was that the composition of the panel was made available by the CA Civ Listing Officer. Note that, under this advance listing system, the outcome of a judge's review could in principle influence his ticketing status but only with *at least* one month's delay. We draw on this observation when arguing that the chance of a change in a judge's ticketing status within 10 days of his review is small.

[14] Information on the timing of coverage in newspaper law reports can be obtained from Westlaw UK.

3 Empirical Strategy

In this Section, we set out a (Rubin Casual) model, state an identification result, and then explain how this result can be used to explore the hypotheses discussed in the Introduction. We conclude by noting how the key identifying assumption can be assessed.

3.1 The Model

We study N panels, indexed by $i = 1, ..., N$, sitting in an appeal court. Each panel is reviewing a judgement taken by a judge sitting in a court of first instance and must decide whether to affirm the judgement or to reverse it (indicating that the judge was wrong on a point of law). The realised outcome for panel i is denoted by Y_i which takes the value 1 if the panel chooses to affirm and 0 if it reverses.

The Assignment Mechanism We normalise the date of each panel's decision to $t = 0$. At other dates, the members of these panels may be sitting alone in a court of first instance, or they may be part of a different team reviewing an unrelated judgment. In the latter case, if a member of panel i is part of the same team as the author of the judgement reviewed at $t = 0$, then we will say that panel i has been *treated*. The date(s) of any such interaction is recorded in a vector of binary treatment status variables, $\mathbf{D_i}$. A typical element of this vector is denoted by $D_{i,t}$ which takes the value 1 if, at date t, a member of panel i sits with the author reviewed at time $t = 0$, and 0 otherwise. To economise on notation, we abstract from observables and assume that $D_{i,t}$ is determined by an unobserved binary selection variable $Z_{i,t}$ (e.g. ticketing status) and chance.[15] In Section 3.2 below, we will make use of the following two assumptions on the distribution of these random variables.

Assumption 1. *Stationary propensity score function.* For all $s > t \neq 0$,

$$Pr[D_{i,t} = 1|Z_{i,t} = 1] = Pr[D_{i,s} = 1|Z_{i,s} = 1] = p < 1$$
$$Pr[D_{i,t} = 1|Z_{i,t} = 0] = Pr[D_{i,s} = 1|Z_{i,s} = 0] = q < p.$$

This assumption states that, if the realisation of the selection variable is the same at two dates t and s, then the probability of treatment will be the same at these two dates. It will hold if the same device is used to randomise conditional on the selection variable. This claim is justified in our setting because

[15] The model can easily be extended to allow for a vector of selection variables, thereby enabling us to incorporate factors such as the number of reviewers, the speed with which the reviewers handle their cases, etc.

the process that randomly matches ticketed judges −the cab rank principle− is applied in the same fashion every period.

Assumption 2. *Markov selection process.* For all $s > t \neq 0$,

$$Pr[Z_{i,t} = 1, Z_{i,s} = 0] = Pr[Z_{i,t} = 0, Z_{i,s} = 1] = \frac{f(s-t)}{2}$$
$$Pr[Z_{i,t} = 1, Z_{i,s} = 1] = Pr[Z_{i,t} = 0, Z_{i,s} = 0] = \frac{1-f(s-t)}{2},$$

where $f(s-t)$ is a continuous increasing function with $\lim_{s-t \to 0} f(s-t) = 0$.

This assumption states that the likelihood of the selection variable taking a different value at two dates t and s is increasing in the elapsed time $s-t$ and, moreover, that the two types of transition (e.g. from ticketed to unticketed and vice versa) are equally likely. It will hold if the assignment mechanism is a Markov process with a symmetric transition rate matrix.[16]

Potential Outcomes Following standard notation, the potential outcome $Y_i(\mathbf{D}_i)$ is the outcome that would be realised if panel i received the treatment profile $\mathbf{D}_i$, and $Y_i(\tilde{\mathbf{D}}_i)$ is the outcome that would be realised if panel i received some different treatment profile $\tilde{\mathbf{D}}_i$. The unit causal effect is therefore $Y_i(\mathbf{D}_i) - Y_i(\tilde{\mathbf{D}}_i)$. Much of the treatment effects literature focuses on the unconditional expectation of unit causal effects (the population average treatment effect). In the next subsection we show that, while it is not possible to identify the population average for any unit causal effect, it is possible to identify the average of a particular unit causal effect for a particular subpopulation.

3.2 Identification

To ease notation, for the remainder of this section we assume that treatment is possible at just two dates: t and s. It suffices to focus on three (of the resulting six) unit causal effects. The first is the effect of treatment at t (a level effect)

$$Y_i(D_{i,t} = 1, D_{i,s} = 0) - Y_i(D_{i,t} = 0, D_{i,s} = 0), \tag{1}$$

[16] This claim is not unreasonable in our setting. If a judge is ticketed today there is a small chance that he will not be ticketed tomorrow due to retirement or a fall in demand in the Court of Appeal; if a judge is not ticketed today there is an equally small chance that he will be ticketed tomorrow due to a promotion or a rise in demand in the Court of Appeal. Empirically, we find that the number of judges who are automatically ticketed stays broadly constant over time. This is consistent with a symmetric transition rate matrix.

the second is the effect of treatment at s (another level effect)

$$Y_i(D_{i,t} = 0, D_{i,s} = 1) - Y_i(D_{i,t} = 0, D_{i,s} = 0), \tag{2}$$

and the third is the difference between (1) and (2) (an order effect)

$$Y_i(D_{i,t} = 1, D_{i,s} = 0) - Y_i(D_{i,t} = 0, D_{i,s} = 1). \tag{3}$$

Our claim is that, although it is not possible to identify a statistic of the distribution of the level effects, it *is* possible to identify a statistic of the distribution of the difference between them, namely the average effect of treatment order for units treated once. More formally, defining this statistic as

$$Order_{t,s} \equiv E[Y_i(D_{i,t} = 1, D_{i,s} = 0) - Y_i(D_{i,t} = 0, D_{i,s} = 1)|D_{i,t} + D_{i,s} = 1]$$

we can state the following result.

Proposition 1. *Under Assumptions 1 and 2,*

$$E[Y_i|D_{i,t} = 1, D_{i,s} = 0] - E[Y_i|D_{i,t} = 0, D_{i,s} = 1] = Order_{t,s} + \Delta(s-t), \text{ with } \lim_{s-t\rightarrow 0} \Delta(s-t) = 0.$$

Proposition 1 states that an estimable quantity is equal to a statistic of the distribution of the unit causal effect in (3) plus a bias term that vanishes as $s - t$ becomes small. We provide a formal proof of this identification result in the Appendix. To see the intuition, note that the bias term will be positive if the increase in $E[Y_i(D_{i,t} = 1, D_{i,s} = 0)]$ from conditioning on $D_{i,t} = 1, D_{i,s} = 0$ rather than $D_{i,t} + D_{i,s} = 1$ is greater than the increase in $E[Y_i(D_{i,t} = 0, D_{i,s} = 1)]$ from conditioning on $D_{i,t} = 0, D_{i,s} = 1$ rather than $D_{i,t} + D_{i,s} = 1$. Since the potential outcomes and treatment variables are orthogonal conditional on unobservables, this can only occur if conditioning on one order rather than another increases the likelihood of a particular realisation of unobservables and this realisation is associated with a higher potential outcome. Given the symmetry imposed by Assumptions 1 and 2, conditioning on one order rather than another has no impact on the likelihood that $Z_{i,t} = Z_{i,s} = 1$ or the likelihood that $Z_{i,t} = Z_{i,s} = 0$. True, conditioning on $D_{i,t} = 1, D_{i,s} = 0$ rather than $D_{i,t} + D_{i,s} = 1$ increases the likelihood that $Z_{i,t} = 1, Z_{i,s} = 0$ and decreases the likelihood that $Z_{i,t} = 0, Z_{i,s} = 1$. However, as $s - t$ becomes small, the likelihood that $Z_{i,t} \neq Z_{i,s}$ (and hence the magnitude of any bias) vanishes. Thus, as $s - t$ becomes small, unobservables will almost certainly stay fixed at either $Z_{i,t} = Z_{i,s} = 1$ or $Z_{i,t} = Z_{i,s} = 0$ and, since the order of treatment is not associated with the relative likelihood of these events, the bias term tends to zero.[17]

3.3 Backward and Forward-Looking Favouritism

To illustrate how Proposition 1 can be used to explore the hypotheses discussed in the Introduction, suppose that $t = -1$ and $s = 1$. One reason that on-the-job interactions could affect the appeal process is what we call *backward-looking favouritism*, i.e. the propensity to look more favourably on the work of judges with whom one has just interacted. Another (not mutually exclusive) channel is what we term *forward-looking favouritism*, i.e. the tendency to be more lenient on the work of judges with whom one is about to interact. Clearly, under backward-looking favouritism we should find that the effect in (1) is positive while under forward-looking favouritism the effect in (2) should be negative.

Now consider a comparison of the units treated before (but not after) the review with the units treated after (but not before) the review. Applying Proposition 1, this difference is approximately equal to $Order_{t,s}$, the average of the unit causal effect in (3) for the subpopulation treated once. A finding that $Order_{t,s} \neq 0$ should then lead us to reject the hypothesis that the appeal process is unaffected by on-the-job interactions. $Order_{t,s} > 0$ would be consistent with backward-looking favouritism, while $Order_{t,s} < 0$ would suggest that forward-looking favouritism is the predominant force.[18] The tests summarised in Section 4 below are based on these observations, although, of course, we allow for observable selection variables and the possibility that treatment can occur on more than two dates.

3.4 Assessing Unconfoundedness

Our empirical strategy rests on the claim that, under Assumption 1 and 2 and for sufficiently small $s - t$, the potential outcomes and treatment variables are unconfounded conditional on possible orders of treatment. Although the validity of this claim cannot be assessed directly (Imbens and Wooldridge 2009), it can be assessed indirectly. Under Assumptions 1 and 2 and for sufficiently small $s - t$, the realisation of an *order* of treatment is not associated with the realisation of the selection variables but is instead determined by chance. Consequently, we should observe: (i) equal proportions of orders of treatment; and (ii) balanced observables across the group treated on the date before (but not after) the review and the group treated on the date after (but not before) the review. The results of these tests are presented in Table 2 Panel B and Table 3.

[17] It is not possible to identify a statistic of the distribution of the unit causal effects in (1) and (2), even as $s-t$ becomes small, because the level of treatment *is* associated with the relative likelihood that $Z_{i,t} = Z_{i,s} = 1$ or $Z_{i,t} = Z_{i,s} = 0$.

[18] Note that this interpretation of $Order_{t,s} < 0$ abstracts from the possibility of backward-looking *antagonism*. This assumption (that interaction effects are solely due to favouritism) is motivated by the literature cited in the Introduction.

4 Data and Estimation

4.1 Sample and Variable Construction

The (realised) outcome variable is straightforward to construct. Using the Westlaw U.K. database of cases in the English superior courts, we are able to link 2298 rulings in the High Court to a corresponding review in the CA Civ. Dropping 36 reviews that occur outside of the working legal year, this gives us a cross-section of 2262 review decisions. The earliest ruling in the High Court is given on 1 February 1980 and the latest review in the CA Civ on 29 November 2005.[19]

Construction of the treatment variables is more complex. An initial consideration is that the Westlaw U.K. database lists when a panel hearing a case finishes its deliberations and hands down its judgment but not when these deliberations start. Fortunately, as noted in Section 2, data from HM Courts and Tribunals Service indicate that reviews in the CA Civ typically last only one or two days. Our response to this missing data problem is to assume that deliberations start and end on the same day; i.e., a panel forms, its members interact, and then the panel dissolves all on the same day.

This discussion suggests that it should be possible to use a *daily* time index. The upside of a daily index is that the elapsed time $s - t$ between $t = -1$ and $s = 1$ is small. The downside is that the size of the treated sample is also small. In fact, just 6 of the 2262 panels in our dataset are treated on the day before and/or the day after the review. Thus, much as sample size concerns force researchers using regression discontinuity design to include observations in windows either side of the selection threshold, we are forced to expand the time index for our treatment indicators beyond a single day.

We proceed by constructing a sample where the length of the time interval is set at 10 days. Table 2 Panel A illustrates the associated sample size (along with a sample where the interval is set at 40 days for comparison purposes). Column 1 shows that 34 panels are treated exactly once, and 41 panels are treated at least once, in total over the 10 days before and the 10 days after the review. Naturally, far more panels are treated when the length of the interval is expanded, as Column 2 confirms. In the remainder of the paper, we use the sample where the length of the time interval is set at 10 days, unless otherwise indicated.

[19] We focus on the final outcome of the case, rather than on individual opinions, because a dissenting opinion features in less than 2 percent of the 15083 CA Civ cases in our database (see Table 1). Westlaw codes each review decision as an affirmance, an affirmance-in-part, a reversal, or a reversal-in-part. Since the number of 'in part' decisions is small, we combine the first two categories, and also collapse the last two categories.

4.2 Comparison of Means

Let $Before_i^{-10/0}$ (respectively $After_i^{0/10}$) denote the number of days in the 10-day period immediately before (respectively after) panel i's decision, upon which there is a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers. Assuming that we have a random sample on Y_i, $Before_i^{-10/0}$ and $After_i^{0/10}$ from the population with $Before_i^{-10/0} + After_i^{0/10} = 1$, the treatment effect *Order* can be consistently estimated by a simple regression of the realised outcome Y_i on a constant and $Before_i^{-10/0}$.[20] The results of this exercise,

$$\widehat{Order} = E[Y_i|Before_i^{-10/0} = 1, After_i^{0/10} = 0] - E[Y_i|Before_i^{-10/0} = 0, After_i^{0/10} = 1], \qquad (4)$$

can be obtained simply by observing the raw data and are reported in Table 4 Panel A.

4.3 Regression Models

We control for observable selection variables by estimating a regression model using the full sample. Specifically, letting $\mathbf{X}_i$ denote a vector of observable characteristics for panel i, we estimate:

$$Y_i = \alpha + \beta \cdot Before_i^{-10/0} + \gamma \cdot \left(Before_i^{-10/0} + After_i^{0/10}\right) + \zeta' \mathbf{X}_i + \varepsilon_i \qquad (5)$$

for $i = 1, ..., 2262$. The model in (5) imposes two additional assumptions, namely that the effect of treatment is linear and (more restrictively) is constant across i. Under these assumptions, $\beta = Order$ and so the OLS estimate $\widehat{\beta}$ provides a robustness check for (4). The results of this estimation exercise are presented in Columns 1 and 2 of Table 5. To control for the possibility of treatment in other periods, we also estimate:

$$Y_i = \theta + \sum_{t=0,10,20,30} \beta_t \cdot Before_i^{(-t-10)/-t} + \sum_{t=0,10,20} \gamma_t \cdot After_i^{t/(t+10)} + \phi \cdot Total_i + \xi' \mathbf{X}_i + \epsilon_i \qquad (6)$$

for $i = 1, ..., 2262$. $Total_i$ is defined as the total number of treatments taking place between 40 days before the review and 40 days after the review, i.e. $Total_i = \sum_{t=0,10,20,30} (Before_i^{(-t-10)/-t} + After_i^{t/(t+10)})$. Since $After_i^{30/40}$ is omitted, $\widehat{\beta}_0$ and $\widehat{\gamma}_0$ are essentially robustness checks for (9) and (10). The results of this exercise are presented in Figure 1.

[20] Note that, in the single treatment sample, the count variables coincide with the binary treatment indicators and (hence) $Before_i^{-10/0} = 1$ implies $After_i^{0/10} = 0$ and $Before_i^{-10/0} = 0$ implies $After_i^{0/10} = 1$.

5 Results

We now summarise our results, postponing any interpretation until Section 5.5.

5.1 Assessing Unconfoundness

Table 2 Panel B reports the tests for equal proportions of orders of treatment using the single treatment samples. Column 1 shows that the proportion of panels treated once in the 10 days before (but not after) the review is lower than the proportion treated once in the 10 days after (but not before) the review. However, this difference, -0.118, is not significantly different from zero at standard inference levels. In Column 2 we find qualitatively similar results if the length of time is set at 40 days. Since the single treatment samples are small, we also test for equality of treatment means using the larger 'any treatment' samples. The differences in treatment means are 0.073 and -0.165 (for the 10 and 40 time intervals respectively) but, again, in every case $t-$tests fail to reject the null hypothesis of equality.

Table 3 reports the results of balancing tests using our main sample where the length of the time interval is set to 10 days.[21] Our primary interest lies in the rank of the reviewed judge and the number of reviewers because (as noted in Section 2) these variables are likely to be associated with both the assignment of treatment and the outcome of the review. Differences in the average of these candidate selection variables across groups would be a particular cause for concern.

We begin by considering the *level* of treatment, since this neatly illustrates that the empirical strategy set out in Section 3 is actually necessary to solve a selection problem. Comparison groups are defined on the basis of $Before_i^{-10/0}$. As expected, there is a large and statistically significant difference in the rank of the reviewed judge across the treated ($Before_i^{-10/0} > 0$) and untreated ($Before_i^{-10/0} = 0$) groups. For instance, 77 percent of the treated group review a judge who holds a post above the rank of Justice the day before the decision. In contrast, just 9 percent of the untreated group review a judge holding such a rank. A $t-$test rejects the null hypothesis of equal means at 1 percent. The difference in the average number of reviewers is far smaller, and the null of equality cannot be rejected at standard levels. Of the other observables, there are statistically significant differences in the coverage of the first instance judgement in newspaper law reports, and the existence of social ties between the reviewed judge and one or more of his reviewers.

In Columns 5-8, we move on to consider the *order* of treatment and define comparison groups on the basis of $Before_i^{-10/0}$ and $After_i^{0/10}$. The difference in the rank of the reviewed judge across the treated

[21] The balancing test results are qualitatively similar when we set the length of the time interval to 40 days. In this sample, only the coverage of the first instance judgement in The Times law report differs across groups.

before but not after group and the treated after but not before group is now far smaller. For instance, 67 percent of the group treated before but not after review a judge who holds a post above the rank of Justice at the decision date, while the corresponding figure for the group treated after but not before is 58 percent. The difference of 9 percentage points is not significant at conventional levels, although this is not particularly informative given the small sample size. A comparison of the *normalised difference* in means gives a more meaningful sense of balance improvement.[22] Reassuringly, the normalised difference drops substantially, from 1.289 in Column 4 to 0.125 in Column 8. The average number of reviewers is exactly 3 for both groups. Only the coverage of the first instance judgement in The Times law report and the existence of social ties differ across groups. For all other observables, the normalised difference falls moving from Column 4 to 8, and lies below the rule of thumb of a quarter.

5.2 Comparison of Means

Panel A in Table 4 shows that the mean affirmance rate for the group treated once in the 10 days before, but not in the 10 days after, the review is 0.533. In stark contrast, the same variable for the group treated once in the 10 days after, but not in the 10 days before, the review is 0.895. The difference, $\widehat{Order} = -0.362$, is statistically significant at the 5 percent level.

5.3 Robustness Checks using Regression Models

The first two columns in Table 5 report estimates of the effect of an additional treatment taking place in the 10 days before the review rather than in the 10 days after the review. Naturally, since few observations are treated more than once, the estimated marginal effect in Column 1 (without controls), -0.334, is similar to the estimated treatment effect $\widehat{Order}$. A $t-$test rejects the null of a zero effect at 1 percent. Column 2 is based on the specification in (5) and controls for the candidate selection variables, as well as other observable characteristics. The estimated effect barely changes and remains significant at 1 percent, indicating that conditioning on observables does little to change the key baseline result. Indeed, the coefficient is stable despite the fact that many of these controls (including the candidate selection variables) are strong predictors of affirmance.

Figure 1 is based on the specification in (6) and reports estimates of the effect of an additional treatment taking place in other time periods relative to the period $After_i^{30/40}$. $\hat{\gamma}_0$ is significantly different from zero, indicating that receipt of an additional treatment *rather than a placebo* is associated with

[22] The normalised difference is equal to the difference in the mean of the covariate between the two groups divided by the square root of the sum of sample variance of the covariate in the two groups. See Imbens and Rubin, forthcoming.

an increase in the probability of affirmance only if it occurs in the 10 days immediately after a review. Importantly, our main result is confirmed even after we control for treatments in other time periods.

5.4 Investigating the Level Effects

The key insight in our identification strategy is that, under Assumptions 1 and 2, conditioning on one order of treatment rather than another has no impact on the likelihood that selection variables are fixed at one level rather than another and so these selection variables can safely be ignored when estimating the effect of treatment order. This order approach is a credible way to investigate the backward and forward-looking favouritism hypotheses but does provide not a definitive answer to our research question; i.e. $\widehat{Order} < 0$ does not actually refute the possibility of backward-looking favouritism, it merely suggests that forward-looking favouritism is the predominant force. It is therefore of interest to investigate the underlying level effects directly.

Although we do not have a clean identification strategy for the level effects, there are two complementary empirical strategies that can help to shed light on the magnitude of these effects. The first is a level approach with a proxy to control for ticketing status (what we term the backward and forward-looking *level tests*), and the second is an order approach based on unanticipated future interaction (the backward and forward-looking *placebo tests*). For reasons discussed below, the estimated treatment effects may be biased upwards. However, under the plausible assumption that this bias is similar across the backward and forward-looking variants of each test, it is still possible to make progress.[23]

Level tests. Let $OtherBefore_i^{-10/0}$ denote the number of days in the 10-day period immediately before panel i's decision, upon which there is a CA Civ judgment where the panel contains the reviewed judge but *not* one of his reviewers. Since $OtherBefore_i^{-10/0} > 0$ implies that the judge reviewed by panel i must be ticketed at the time of the review decision, we can attempt to estimate the level effect of a recent on-the-job interaction via the following comparison of means

$$\begin{aligned} Level\widehat{Backward} = {} & E[Y_i | Before_i^{-10/0} = 1, After_i^{0/10} = 0] \\ & - E[Y_i | Before_i^{-10/0} = 0, After_i^{0/10} = 0, OtherBefore_i^{-10/0} = 1]. \end{aligned} \tag{7}$$

All panels used in this comparison are reviewing a judge who is currently ticketed but only those for whom $Before_i^{-10/0} = 1$ will have experienced an on-the-job interaction with this judge shortly before

[23] We are grateful to an anonymous referee for encouraging us to pursue this line of reasoning.

the review decision. Similarly, letting $OtherAfter_i^{0/10}$ denote the number of days in the 10-day period immediately after panel i's decision, upon which there is a CA Civ judgment where the panel contains the reviewed judge but not one of his reviewers, we can attempt to estimate the level effect of a future anticipated on-the-job interaction via the comparison

$$\begin{aligned} \widehat{LevelForward} = {} & E[Y_i|Before_i^{-10/0} = 0, After_i^{0/10} = 1] \\ & - E[Y_i|Before_i^{-10/0} = 0, After_i^{0/10} = 0, OtherAfter_i^{0/10} = 1]. \end{aligned} \tag{8}$$

We hesitate to claim that the level effects are formally identified under this strategy, however. Although ticketing status is held constant, it does not follow that *perceived* quality is held constant since the untreated panels may not be aware that their reviewed judge is ticketed.

Placebo tests. Our second strategy returns to the order approach but exploits the fact that, at the time of the review decision, panel members are unlikely to know (or even if they know, unlikely to give much weight to) who they will be working with at hearings taking place more than a month in the future. Let $After_i^{30/40}$ denote the number of days in the 10-day period starting 30 days after the review. Since unanticipated future interaction should have no causal effect, we can attempt to estimate the level effect of a recent on-the-job interaction via the comparison

$$\widehat{PlaceboBackward} = E[Y_i|Before_i^{-10/0} = 1, After_i^{30/40} = 0] - E[Y_i|Before_i^{-10/0} = 0, After_i^{30/40} = 1], \tag{9}$$

and the level effect of an anticipated future on-the-job interaction via the comparison

$$\widehat{PlaceboForward} = E[Y_i|After_i^{0/10} = 1, After_i^{30/40} = 0] - E[Y_i|After_i^{0/10} = 0, After_i^{30/40} = 1]. \tag{10}$$

Again a note of caution needed: since the elapsed time $s - t$ is longer than in the test proposed in Section 3, it is less plausible that unobservables will be held fixed. In particular, panels treated with a 'placebo' may not be aware that their reviewed judge is ticketed, raising the possibility of an upward perceived quality bias just as in the level approach described above.

Results. Table 4 Panels B and C show that $\widehat{LevelBackward} = -0.019$ is not significantly different from zero at standard levels, but that $\widehat{LevelForward} = 0.346$ is positive and significant at the 1 percent level. Panels D and E confirm that $\widehat{PlaceboBackward} = -0.012$ is not significantly different from zero at standard inference levels, while $\widehat{PlaceboForward} = 0.274$ is positive and statistically significant at the

10 percent level. For completeness, we also apply the level approach to our regression models using the full sample. Column 3 of Table 5 shows that, in the absence of covariates, the marginal (level) effect of an additional treatment taking place in the 10 days before the review is not significantly different from zero. Adding a full set of controls and $OtherBefore_i^{-10/0}$ as a proxy for ticketing status[24] in Column 4 does not change this conclusion. Column 5 reports the marginal (level) effect of an additional treatment taking place in the 10 days after the review. This estimate is positive and strongly significant, and remains so in Column 6 when we add controls and $OtherAfter_i^{0/10}$ as a proxy for ticketing status.

5.5 Interpretation

We interpret the finding that $\widehat{Order}$ is significantly different from zero as evidence that the appeals process is not above the influence of on-the-job interaction. An alternative interpretation is that this difference in mean affirmance rates is due to selection bias. Various pieces of evidence suggest that this is unlikely. The tests for equal proportions of orders of treatment and balanced observables are consistent with unconfoundedness of potential outcomes and treatment variables conditional on possible orders of treatment; a claim that is further substantiated by the fact that, in our order regression models, controlling for observables has little effect.

Our estimate of *Order* is not just statistically significant but also large in magnitude, particularly since typically only one of the three reviewers experiences an on-the-job with the reviewed judge. The finding that $\widehat{Order}$ is negative is consistent with the predominant force being forward-looking favouritism motivated by a fear of awkwardness and/or reprisal in the future on-the-job interaction.

The results from Section 5.4 indicate that forward-looking favouritism may actually be the only force at work. Recall that we are unable to reject the null hypotheses that $Level\widehat{Backward}$ and $Placebo\widehat{Backward}$ are zero. In particular, the point estimates are close to zero, suggesting that both backward-looking favortism and any bias due to unobserved perceptions of quality must be small. On the other hand, both $Level\widehat{Forward}$ and $Placebo\widehat{Forward}$ are positive and significant. Since the perceived quality bias should be similar for these forward-looking tests as for the backward-looking tests (i.e. minimal), these positive and significant estimates can therefore be taken as evidence of forward-looking favouritism, rather than simply bias.

[24] Note that, in the full sample, this can only be a rough proxy for ticketing status because the absence of a CA Civ judgement where the panel contains the reviewed judge during a given a time period does not imply that this judge is not ticketed during that time period.

6 Exploring the Forward-looking Favouritism Mechanism

Having argued that forward-looking favouritism appears to be the only force at work, we now assess this mechanism in more detail.

6.1 The Rationale for Forward-looking Favouritism

We begin by looking for evidence to substantiate the rationale for forward-looking favouritism, namely that panel members will anticipate that an affirmance makes it easier to work alongside the reviewed judge immediately after the review. To do so, we compare working relationships in on-the-job interactions that occur shortly before the review with those in on-the-job interactions that occur shortly after the review. Although many aspects of on-the-job interactions are beyond measurement, it is possible to gain an insight into these working relationships by looking for the presence of *dissenting opinions*. Panels sitting in the CA Civ are not required to reach unanimous agreement and a panel member who finds himself in the minority can signal this fact by publishing his own dissenting opinion. Such behaviour is widely deemed to be uncollegial (Cross and Tiller 2008) and, in the English system at least, is rare (see Table 1).

Since dissents are rare events, we expand the sample size by studying time intervals of 40 days. We find that, for the group treated once in the 40 days before the review but not in the 40 days immediately after the review, the mean dissent rate in the on-the-job interaction is 7 percent. For the group treated once in the 40 days immediately after the review but not in the 40 days before the review, the mean dissent rate in the on-the-job interaction is lower, at 3 percent. Disaggregating by the type of review decision, we find that the dissent rate is lower when the on-the-job-interaction occurs after rather than before an affirmance (4 percent versus 10 percent). This is consistent with a rationale for forward-looking favouritism. However, there is no evidence that the dissent rate is higher when the on-the-job interaction occurs after rather than before a reversal (the dissent rate is zero for both groups). Power is an obvious concern here and, unsurprisingly, these differences are not statistically significant. As such, we view these results as merely suggestive of a rationale for forward-looking favouritism.

6.2 Heterogeneity in Forward-looking Favouritism

If forward-looking favouritism really is the mechanism at work in Tables 4 and 5, one might expect the size of the treatment effect to vary with the nature of the pre-existing relationship between the reviewers and the reviewed judge. For instance, a reviewer who is already socially connected to the reviewed judge might be more prone to forward-looking favouritism because it is particularly awkward to work alongside a ‘friend’ immediately after reversing one of his judgements. On the other hand, one might expect that reviewers who are more senior than their reviewed judge to be less prone to forward-looking favouritism, either because there is less stigma when a junior is reversed by a senior or because there is less scope for future reprisals.

Although we have data on the educational and social networks of the judges in our sample, there are too few instances of a tie between a reviewer and reviewed judge to test for heterogeneity in the treatment effect along this dimension. As Table 3 indicates, none of the reviewers that were connected to their reviewed judge via an on-the-job interaction in the 10 days before or after the review were also at school or university together with this judge and only 5 (3) worked at the same legal chambers (share the same social club) as this judge. We can, however, look for heterogeneity along the seniority dimension. Since our objective is to explore the effects of forward-looking favouritism, we focus on groups that are treated with a single interaction *after* the review. In this sample, all of the reviewers that experience this on-the-job interaction hold the rank of Lord Justice at the time of the review (with one exception who is a Law Lord). In contrast, only 63 percent of the reviewed judges hold this rank or above at the time of the review. Since there are no observations where the reviewer is less senior than the reviewed judge, we split the observations into ‘more senior’ and ‘same rank’ subsamples.

Our results are presented in Table 6. Panel A shows that the mean affirmance rate for the group where a reviewer anticipates an imminent on-the-job interaction with a reviewed judge who is less senior than himself is 0.667. The mean affirmance rate for the group where a reviewer experiences an *unanticipated* on-the-job interaction with a reviewed judge who is less senior than himself is 0.750. The difference in means, −0.083, is both economically and statistically insignificant, indicating that we have failed to find evidence of a forward favouritism effect in this ‘more senior’ subsample. Panel B shows that the mean affirmance rate for the group where a reviewer anticipates an imminent on-the-job interaction with a reviewed judge who holds the same rank as himself is 1.000. The mean affirmance rate for the group where a reviewer experiences an *unanticipated* on-the-job interaction with a reviewed judge who holds the same rank as himself is 0.500. The difference in means, 0.500, is larger than in the full sample and is statistically different from zero at 5 percent, while the difference-in-difference estimate for these subsamples is statistically different from zero at 10 percent. It follows that reviewers do indeed suffer less from forward-looking favouritism bias when assessing junior colleagues than when assessing peers of the same rank. As we note in the Conclusion, this finding cautions against the trend towards decentralised open performance appraisals, and highlights the need for anonymity in ‘360 degree’ reviews.

6.3 Further Consequences of Forward-looking Favouritism

In this section, we present a simple theoretical framework that enables us to draw out, and then test, additional empirical predictions.

Set-up Our starting assumption is that there is a correct ruling, a "state of the world" $x \in \{0, 1\}$. For concreteness, we let $x = 0$ denote the state where the reviewed judged was right and should be affirmed, and $x = 1$ the state where the reviewed judge was wrong and should be reversed. Reflecting aggregate affirmance rates, the panel's prior belief that $x = 0$ is denoted by $\mu > 1/2$. The panel cannot observe x but can combine its own legal knowledge with the facts of the case to revise its prior belief. We equate this process with the generation of an informative private signal on x, $s \in \{0, 1\}$. The precision of this signal is a binary random variable that takes a high realisation $p = p_H$, and a low realisation $p = p_L$, with equal probability. The panel also receives a second (orthogonal) signal, $\sigma \in \{0, 1\}$, indicating whether a reviewer will work alongside the reviewed judge after the review. Having observed p, s and σ, the panel makes a ruling $r \in \{0, 1\}$ affirming or reversing the reviewed judge. It will be helpful to define $\gamma_{p,s,r}$ as the belief of a panel with precision p and signal s that this ruling r is correct.

After the panel has made its ruling, the parties to the case may lodge a legal challenge to the House of Lords. Rather than modeling this behaviour explicitly, we assume that the panel expects to see a legal challenge if its decision is incorrect (fails to match x).[25] The panel then disbands and, if $\sigma = 1$, a panel member works alongside the reviewed judge.

The panel incurs disutility from two sources: damage D if the decision produces a legal challenge, and cost C if the decision is a reversal *and* a reviewer subsequently works alongside the reviewed judge. To make concrete predictions, we place the following restriction on parameter values.

Assumption 3. The parameters satisfy the following inequalities:

$$p_H > \frac{(C + D)\mu}{C(2\mu - 1) + D} > p_L > \mu. \tag{11}$$

To summarise, the timing runs as follows. The panel learns the precision p and realisation s of its signal on x, and the realisation of its signal on forthcoming on-the-job interactions σ, and then makes its review decision r. A legal challenge is lodged with probability $\gamma_{p,s,r}$ and, if $\sigma = 1$, a reviewer works alongside the reviewed judge. Finally, the panel's payoff is realised. It follows that the panel chooses r to maximise its expected payoff: $-(1 - \gamma_{p,s,r}) \cdot D - 1 [r = 1, \sigma = 1] \cdot C$.

[25] This is a simple way to capture the intuitive idea (discussed in Blanes i Vidal and Leaver 2013) that the panel will perceive the likelihood of a legal challenge to be lower when it is more confident that its ruling is correct.

Analysis and Predictions Consider a panel with signals $s = \sigma = 1$. Since $p_H > p_L > \mu$, this panel believes that a reversal is more likely to be correct than an affirmance. To maximise the probability of a correct decision this panel should reverse the reviewed judge. Consequently, we will say that there is a *forward-looking favouritism bias* in decision-making if, for either realisation of p, this panel affirms the reviewed judge.

When deciding on a ruling, this panel considers both the likelihood of a legal challenge and the (extra-legal) cost of reversing the reviewed judge. This panel reverses only if the payoff from doing so $-(1-\gamma_{p,1,1}) \cdot D - C$ is no smaller than the payoff from affirming, namely $-(1-\gamma_{p,1,0}) \cdot D$ or, equivalently, only if $(\gamma_{p,1,1} - \gamma_{p,1,0}) \cdot D \geq C$. Applying Bayes' rule to establish $\gamma_{p,1,1} - \gamma_{p,1,0} = (p-\mu)/(p+\mu-2p\mu)$ and rearranging for p, this necessary condition for a reversal can be written as $p \geq (C+D)\mu / \left[C(2\mu - 1) + D\right]$. Given Assumption 3, it follows that the reviewed judge is reversed if $p = p_H$ but affirmed if $p = p_L$.

Now consider a panel with signals $s = 0, \sigma = 1$. To maximise the probability of making a correct decision, this panel should affirm the reviewed judge. Since this ruling avoids the extra-legal cost of reversing, decision-making is unbiased. Similarly, when $\sigma = 0$, the panel has no (extra legal) reason to fear a reversal and so decision-making is unbiased for both realisations of s. These observations enable us to state the following result.

Proposition 2. *An anticipated on-the-job interaction causes a forward-looking favouritism bias in decision-making that:*

i. *increases the probability that the panel affirms,* $\Pr\left[r = 0 | \sigma = 1\right] > \Pr\left[r = 0, \sigma = 0\right]$*;*

ii. *increases the probability that the review decision is incorrect,* $\Pr\left[r \neq x | \sigma = 1\right] > \Pr\left[r \neq x | \sigma = 0\right]$*;*

iii. *increases the probability that an affirmance is incorrect,*
$\Pr\left[x = 1 | r = 0, \sigma = 1\right] > \Pr\left[x = 1 | r = 0, \sigma = 0\right]$*; but*

iv. decreases *the probability that a reversal is incorrect,* $\Pr\left[x = 0 | r = 1, \sigma = 1\right] > \Pr\left[x = 0 | r = 1, \sigma = 0\right]$.

There is a forward-looking favouritism bias in decision-making because, when $p = p_L$ and $s = \sigma = 1$, the panel is insufficiently confident that reversing the reviewed judge is the correct decision. As a result, the extra-legal cost of reversing outweighs the expected (legal) benefit and the panel affirms. It follows that, averaging over realisations of p and s, the probability that the panel affirms the reviewed judge conditional on an anticipated on-the-job interaction is higher than the probability that the panel affirms the reviewed judge conditional on no anticipated on-the-job interaction. This is the prediction that was tested, and confirmed, in Section 5.

It also follows that the probability that the review decision is incorrect conditional on an anticipated on-the-job interaction is higher than the probability that the review decision is incorrect conditional on no anticipated on-the-job interaction. Similarly, the probability that an affirmance is incorrect conditional on an anticipated on-the-job interaction is higher than the probability that an affirmance is incorrect conditional on no anticipated on-the-job interaction. This is because, with positive probability, the panel affirms the reviewed judge to avoid the cost C despite being aware that a reversal is more likely to be the correct decision. In contrast, the probability that a reversal is incorrect conditional on an anticipated on-the-job interaction is *lower* than the probability that a reversal is incorrect conditional on no anticipated on-the-job interaction. This is because the panel reverses the reviewed judge only if this decision is supported by a highly precise signal. These three predictions are tested below.

Empirical Results To test Proposition 2 Parts ii-iv we require an indicator of whether review decisions are correct. Following the legal literature, we use two different data sources: (i) legal challenges (appeals) to the House of Lords and (ii) citations by judges in other cases. The first measure is consistent with our theoretical framework: a legal challenge should be more likely to occur when the review decision is incorrect than when it is correct. The logic for using judicial citations is that other judges should be less likely to apply the panel's legal reasoning (a positive citation) when the decision is incorrect than when it is correct. Similarly, other judges should be more likely to criticise the panel's legal reasoning (a negative citation) when the review decision is incorrect than when it is correct.[26]

Our results are presented in Figure 2. Since our objective is to explore the effects of forward-looking favouritism, we again focus on groups that are treated with a single interaction *after* the review. The headline finding is that there is strong evidence to support Proposition 2 Part iv: a legal challenge of a reversal is *less* likely among the group where the panel is treated with an anticipated interaction than among the group where the panel is treated with an unanticipated interaction. Other results, while consistent with Proposition 2, are statistically insignificant.

Panel A pools across all review decisions. The mean appeal rate for the group treated once in the 10 days immediately after the review but not in the period 30-40 days after the review −i.e. the group treated with an anticipated interaction− is 0.095 (first bar). The mean appeal rate for the group treated in the period 30-40 days after the review but not in the 10 days starting immediately after the review −i.e. the group treated with an unanticipated interaction− is 0.167 (fourth bar). The mean positive citation rate for the group treated with an anticipated interaction is 0.333, as is the mean

[26] For a more detailed description of judicial citations in English courts, see Blanes i Vidal and Leaver (2013). Much like dissenting opinions, negative citations are rare. As Table 1 indicates, just 5 percent of the 15083 CA Civ cases in our database receive a negative citation.

positive citation rate for the group treated with an unanticipated interaction. Thus, for both appeals and positive citations, there is no evidence to support Proposition 2 Part ii. However, the mean negative citation rate for the group treated with an anticipated interaction is 0.095, while the mean negative citation rate for the group treated with an unanticipated interaction is 0. The positive sign of this difference in means is consistent with Proposition 2 Part ii, although the estimate is not statistically significant.

Figure 2 Panel B uses the same observations but disaggregates by the review decision. The mean appeal rate for the group where the panel affirms the reviewed judge and is treated with an anticipated interaction is 0.11 (first bar), while the mean appeal rate for the group where the panel affirms the reviewed judge and is treated with an unanticipated interaction is 0 (absence of a fourth bar). An identical pattern is observed for negative citations. For positive citations, the mean positive citation rate for the group where the panel affirms the reviewed judge and is treated with an anticipated interaction is 0.33 , while the mean positive citation rate for the group where the panel affirms the reviewed judge and is treated with an unanticipated interaction is 0.42. The signs of all three differences in means are consistent with Proposition 2 Part iii, although again the estimates are not statistically significant.

Turning to reversals, the mean appeal rate for the group where the panel reverses the reviewed judge and is treated with an anticipated interaction is 0 (absence of a seventh bar), while the mean appeal rate for the group where the panel reverses the reviewed judge and is treated with an unanticipated interaction is 0.40 (tenth bar). The negative sign of this difference in means is consistent with Proposition 2 Part iv, and the estimate is statistically significant at the 10 percent level ($p = 0.07$). Moreover, the difference in the treatment effect of an anticipated interaction on appeals when the review decision is an affirmance rather than a reversal is positive (0.51) and significant at the 5 percent level. The mean positive citation rate for the group where the panel reverses the reviewed judge and is treated with an anticipated interaction is 0.33, while the mean positive citation rate for the group where the panel reverses the reviewed judge and is treated with an unanticipated interaction is 0.20. The positive sign of this difference in means is consistent with Proposition 2 Part iv, although this estimate is not statistically significant. The absence of a bar in the remaining categories indicates that there is no difference in the mean negative citation rate across groups.

Summing up, there is descriptive evidence to support Proposition 2 Parts ii and iii, and stronger statistically significant evidence to support with Proposition 2 Part iv. These findings substantiate our claim that anticipated on-the-job interaction can introduce a forward-looking favouritism bias into judicial decision-making.

7 Concluding Remarks

Open peer-review, where the identities of the reviewers and reviewees are public, is used to assess performance in a variety of settings, including legislative and judicial branches of government, academia, and professional service firms. Proponents claim that removing the anonymity of reviewers brings ethical and intellectual benefits and, by fostering reputational accountability, could also minimise reviewer bias driven either by discrimination, or favouritism motivated by existing personal ties. However, it has also been noted that open reviewing could, in principle, lead to alternative forms of bias as reviewers, fearing future awkwardness and/or reprisal for their public criticism, take a lenient "kid gloves" approach. This paper uses data from the English superior courts to explore whether open peer-review is subject to bias and, if so, whether the underlying mechanism can be attributed to backward-looking favouritism, forward-looking favouritism, or both.

Our empirical strategy exploits the random timing of on-the-job interaction between reviewers (sitting in the Court of Appeal) and reviewees (who have heard cases in the High Court). The main findings are that reviewers show a reluctance to reverse the judgements of reviewees with whom they are about to interact, and that this effect is stronger when reviewer and reviewee share the same rank. The average bias is substantial: the proportion of reviewers that affirm their reviewee is 30 percentage points higher in the group where reviewers know they will soon work with their reviewee, relative to groups where such interaction occurs before the review, or after the review but is unanticipated. We interpret these findings as evidence that, when lacking the protection of anonymity and when assessing a (true) peer rather than a junior colleague, reviewers may indeed take a lenient "kid gloves" approach.

To explore this mechanism further, we present a model of forward-looking favouritism that yields predictions relating to the *quality* of review decisions. Consistent with these predictions, we find that: reversals taken in advance of an anticipated on-the-job interaction are significantly less likely to result in a legal challenge (to be of low quality) than reversals taken in advance of an unanticipated interaction; and the difference in the effect of an anticipated interaction on the likelihood of a legal challenge when the review decision is an affirmance rather than a reversal is positive and strongly significant.

Taken together, our results suggest that the reversal rate in the Court of Appeal may be inefficiently low. This conclusion is troubling since it cannot be explained away by the argument that judges are human beings and so their personal histories unavoidably shape their legal views. Instead, echoing previous evidence (Cross and Tiller 2008), our paper points toward the existence of a collegial culture in which judges actively avoid public contradiction of their peers.[27]

[27] Related to this, we provide a new rationale for the strong dissent aversion that is often found in appellate panels.

Forward-looking favouritism bias seems less likely under a system of blind review because, when a reviewee is unaware of the identity of his reviewers, reprisals are not possible and the motivation to pre-empt is reduced. With blind review not an option, the main policy implication of our research for the judiciary relates to the listing system. HM Courts and Tribunals Service should consider reforming the listing process to ensure that judges cannot anticipate that they will soon sit with colleagues affected by their decisions. This could be achieved by limiting downward movement of judges (i.e. a Lord Justice hearing a case in the High Court) since this would increase the distance between reviewers and reviewees in the judicial hierarchy and hence lower the probability of an on-the-job interaction shortly after the review.[28] Naturally, the benefit of reduced bias would need to be weighed against the cost of expanding the High Court bench, as well as a potential loss of expertise. More laboriously, the CA Civ Listing Officer could vet potential panels for the presence of a reviewer-reviewee pair (two judges, one of whom will have just reviewed the other) and then reallocate one of these judges before the listings are made public. To the extent that reversal rates could be similarly influenced by non-random, explicitly social interactions (of the type documented in Blanes i Vidal and Leaver 2011), it would also be prudent to exercise caution when using appeal judgements to assess the performance of the High Court Bench (c.f. The Judges' Council 2003).[29]

Turning to the generalisability and wider policy implications of our research, our view is that similar behaviour could be present in judicial settings in other countries. In the U.S., for instance, two of the necessary conditions appear to be met: there is evidence that judges hearing cases in federal district courts are reversal averse (Shepherd 2011); and these judges sometimes work alongside their reviewers following a promotion or a temporary assignment to the Courts of Appeals. Whether on-the-job interaction occurs with a similar frequency to the English Court of Appeal and, in particular, sufficiently close to review decisions to be anticipated by members of the panel is an open question, worthy of future study.

A related issue is whether open peer-review is likely to create a forward-looking favouritism bias

First, if reversal is socially awkward (as required by our forward-looking favouritism story), judges may choose to seek cover through a unanimous opinion. A norm of unanimity may then arise so that the reversal is not necessarily attributed to particular judges. Second, our finding that judges pre-empt the possibility of dissent by being particularly lenient with the cases of judges with whom they are about to interact is itself an explanation for the fact that dissents are relatively rare in appellate panels.

[28] Limiting the upward movement judges (i.e. a Justice hearing a case in the Court of Appeal) would also lower this probability. Since there is no evidence of forward-looking favouritism bias when the reviewed judge holds the rank of Justice at the time of the review this further step may not be necessary.

[29] Caution would be especially warranted if it is difficult to discount any forward-looking favouritism that could have affected the reviews of first instance judgements. This is likely to be the case when the party assessing the performance lacks access to historical data documenting social and/or on-the-job interactions.

in other professional settings. The bias that we identify is certainly consistent with the results from field experiments of open versus single-blind review within performance appraisal systems (c.f. Antonioni 1994, Afonso et al. 2005, and Kagan et al 2006). This literature points to the existence of 'rating inflation' under open peer-review but has tended to focus on lower-level employees and has not commented on the underlying mechanism. Our findings indicate that such bias could also be present among higher-level employees taking 'high stakes' decisions and, moreover, that this behaviour may be driven by reviewers' fears of awkwardness and/or reprisal in imminent face-to-face interaction with their reviewee. The trend in performance appraisal techniques is for reviews to be open (available to the employee), decentralised (conducted by the employee's immediate line manager rather than upper-level management), and to include multi-rater '360 degree' feedback (from customers, subordinates and peers). Table 6 suggests that firms should reconsider the merits of decentralised open performance appraisals, and highlights the need for anonymity in '360 degree' reviews.

Turning to scientific publishing, the results from the small number of randomised controlled trials of open versus single-blind review at medical journals are also consistent with our finding (c.f. van Rooyen et al 1999 and Walsh et al 2000). Our results suggest that further experimentation with open peer-review should proceed with care, and may not be appropriate in every discipline.[30] Indeed, our paper provides quantitative econometric support for the following view expressed to the U.K. Government's Science and Technology Committee during its 2011 investigation into 'Peer Review in Scientific Publications':

> Some editors have said to us "We work in a very narrow field. Everybody knows everybody else. It just would not work to have this open peer review." There are different options. (...) My opinion is that it depends on the discipline. With a discipline as big as medicine, where there are hundreds of thousands of people all around the world you can ask and they probably don't bump into each other the next day, open peer review seems to work. In much narrower and more specialised fields, it perhaps does not, and the traditional system of the blinded review is perhaps better.[31]

References

[1] Afonso, Nelia, Lavoisier Cardozo, Oswald Mascarenhas, Anil Arahnha and Chirag Shah 'Are Anonymous Evaluations a Better Assessment of Faculty Teaching Performance? A Comparative Analysis of Open and Anonymous Evaluation Processes', *Faculty Development*, 2005, 37:1, 43-47.

[30] The extrapolation from appellate reviews to scientific publishing relies on the notion that legal decisions are correct or incorrect, rather than based on different value judgments. It is precisely this former framework that we adopt in Section 6.3 (see, among others, Spitzer and Talley 2000).

[31] Evidence from the chair of the Committee on Publication Ethics (Science and Technology Committee 2011, Para 19).

[2] Antonioni, David 'The Effect of Feedback Accountability on Upward Appraisal Ratings', *Personnel Psychology*, 1994, 47:2, 349-356.

[3] Bailey, S.H., Jane Ching, M.J. Gunn, and David Ormerod *Smith, Bailey and Gunn on the Modern English Legal System*, 2002, London: Sweet and Maxwell.

[4] Blackwell, Michael 'Measuring the Length of the Chancellor's Foot: Quantifying How Legal Outcomes Depend on the Judges Hearing the Case and Whether Such Variation Can be Explained by Characteristics of the Judges', 2011, available at http://ssrn.com/abstract=1855719.

[5] Blanes i Vidal, Jordi and Clare Leaver 'Are Tenured Judges Insulated From Political Pressure?', *Journal of Public Economics*, 2011, 95, 570-586.

[6] Blanes i Vidal, Jordi and Clare Leaver 'Social Interactions and the Content of Legal Opinions', *Journal of Law, Economics, and Organization*, 2013, 29:1, 78-114.

[7] Blank, Rebecca 'The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review', *American Economic Review*, 1991, 81:5, 1041-1067.

[8] Choi, Stephen, Mitu Gulati and Eric Posner 'What do Federal District Judges Want?: An Analysis of Publications, Citations, and Reversals', 2010, available at http://ssrn.com/abstract=1536723.

[9] Cole, Stephen, Jonathan Cole and Gary Simon 'Chance and Consensus in Peer Review', *Science*, 1981, 214, 881-886.

[10] Cross, Frank and Emerson Tiller 'Understanding Collegiality on the Court', *University of Pennsylvania Journal of Constitutional Law*, 2008, 10:2, 257-271.

[11] Epstein, Lee, Andrew Martin, Kevin Quinn and Jeffrey Segal 'Circuit Effects: How the Norm of Federal Judicial Experience Biases the Supreme Court', *University of Pennsylvania Law Review*, 2009, 157, 833-880.

[12] Fabiato, Alexandre 'Anonymity of Reviewers', *Cardiovascular Research*, 1994, 28, 1134-1139.

[13] Fafchamps, Marcel, Sanjeev Goyal and Marco van der Leij 'Matching and Network Effects', *Journal of the European Economic Association*, 2010, 8:1, 203-231.

[14] Gilbert, Julie, Elaine Williams and George Lundberg 'Is There Gender Bias in JAMA's Peer Review Process?' *Journal of the American Medical Association*, 1994, 272, 139-142.

[15] Ginther, Donna, *et al* ‘Race, Ethnicity, and NIH Research Awards’, *Science*, 2011, 333, 1015-1019.

[16] Godlee Fiona ‘Making Reviewers Visible: Openness, Accountability and Credit’, *Journal of the American Medical Association*, 2002, 287:21, 2762-2765.

[17] Godlee Fiona, Catharine Gale and Christopher Martyn ‘Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports: A Randomized Controlled Trial’, *Journal of the American Medical Association*, 1998, 280:3, 237-240.

[18] Griffith, John A. G. ‘The Politics of the Judiciary’, 1997, London: Fontana Press.

[19] Imbens, Guido and Jeffrey Wooldridge ‘Recent Developments in the Econometrics of Program Evaluation’, *Journal of Economic Literature*, 47:1, 5-86.

[20] Imbens, Guido and Donald Rubin *Causal Inference in Statistics and the Social Sciences*, forthcoming, Cambridge and New York: Cambridge University Press.

[21] The Judges’ Council ‘Response to the Consultation Papers on Constitutional Reform’, 2003, mimeo.

[22] Kagan, Ilya, Ronit Kigli-Shemesh and Nilli Tabak ‘Let Me Tell You What I Really Think About You - Evaluating Nursing Managers Using Anonymous Staff Feedback’, *Journal of Nursing Management*, 2006, 14:5, 356-365.

[23] Kassirer, Jerome and Edward Campion ‘Peer Review: Crude and Understudied, but Indispensable’, *Journal of the American Medical Association*, 1994, 272:2, 96-97.

[24] Lee, David S. and Thomas Lemieux ‘Regression Discontinuity Designs in Economics’, *Journal of Economic Literature*, 2010, 48, 281-335.

[25] Li, Danielle ‘Information, Bias, and Efficiency in Expert Evaluation: Evidence from the NIH’, 2012, mimeo Northwestern University.

[26] Moody, James ‘Race, School Integration, and Friendship Segregation in America’, *American Journal of Sociology*, 2001, 107, 679-716.

[27] Murphy, Kevin R. and Jeanette Cleveland *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*, 1995, California: Sage Publications.

[28] Peters, Douglas and Stephen Ceci 'Peer-review Practices of Psychological Journals: The fate of Published Articles, Submitted Again", *Behavioral and Brain Sciences*, 1982, 5:2, 187-255.

[29] Robertson, Peter 'Towards Open Refereeing', *New Scientist*, 1976, 71, 410.

[30] Robertson, David 'Judicial Discretion in the House of Lords', 1998, Oxford: Clarendon Press.

[31] Salzberger, Eli and Paul Fenn 'Judicial Independence: Some Evidence from the English Court of Appeal', *Journal of Law and Economics*, 1999, 42:2, 831-847.

[32] Science and Technology Committee *Eighth Report: Peer Review in Scientific Publications*, 2011, HC 856, available at www.parliament.uk.

[33] Shepherd, Joanna 'Measuring Maximising Judges: Empirical Legal Studies, Public Choice Theory, and Judicial Behavior', 2011, available at http://ssrn.com/abstract=1910918.

[34] Sisk, Gregory, Michael Heise and Andrew Morriss 'Charting the Influences on the Judicial Mind: An Empirical Study of Judicial Reasoning', *New York University Law Review*, 1998, 73:5, 1337-1500.

[35] Smith, Richard 'Opening Up BMJ Peer Review', *British Medical Journal*, 1999, 318, 4-5.

[36] Spitzer, Matt and Eric Talley 'Judicial Auditing', *Journal of Legal Studies*, 2000, 29, 649-683.

[37] Steinbuch, Robert 'An Empirical Analysis of Reversal Rates in the Eighth Circuit During 2008', *Loyola of Los Angeles Law Review*, 2009, 43, 51-65.

[38] Sunstein, Cass *Are Judges Political? An Empirical Analysis of the Federal Judiciary*, 2006, Washington, D.C.: Brookings Institution Press.

[39] van Rooyen, Susan, Fiona Godlee, Stephen Evans, Nick Black and Richard Smith 'Effect of Open Peer Review on Quality of Reviews and on Reviewers' Recommendations: A Randomised Trial', *British Medical Journal*, 1999, 318, 23-27.

[40] Walsh, Elizabeth, Maeve Rooney, Louis Appleby and Greg Wilkinson 'Open Peer Review: a Randomised Controlled Trial', *British Journal of Psychiatry*, 2000, 176:1, 47-51.

[41] Wenneras, Christine and Agnes Wold 'Nepotism and Sexism in Peer Review', *Nature*, 1997, 387, 341-343.

Appendix

Proof of Proposition 1. Using the definition of $Order_{t,s}$, the bias term, $\Delta(s-t)$, is:

$$\begin{aligned}\Delta(s-t) =& E[Y_i(D_{i,t}=1, D_{i,s}=0)|D_{i,t}=1, D_{i,s}=0] - E[Y_i(D_{i,t}=1, D_{i,s}=0)|D_{i,t}+D_{i,s}=1] - \\ &(E[Y_i(D_{i,t}=0, D_{i,s}=1)|D_{i,t}=0, D_{i,s}=1] - E[Y_i(D_{i,t}=0, D_{i,s}=1)|D_{i,t}+D_{i,s}=1]).\end{aligned}$$

Applying the Law of Total Probability, we can re-write this as:

$$\begin{aligned}&\Delta(s-t) = \sum_{y,w\in\{0,1\}} \\ &(E[Y_i(1,0)|Z_{i,t}=y, Z_{i,s}=w, D_{i,t}=1, D_{i,s}=0] \times Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}=1, D_{i,s}=0] - \\ &\;E[Y_i(1,0)|Z_{i,t}=y, Z_{i,s}=w, D_{i,t}+D_{i,s}=1] \times Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}+D_{i,s}=1] - \\ &(E[Y_i(0,1)|Z_{i,t}=y, Z_{i,s}=w, D_{i,t}=0, D_{i,s}=1] \times Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}=0, D_{i,s}=1] - \\ &\;E[Y_i(0,1)|Z_{i,t}=y, Z_{i,s}=w, D_{i,t}+D_{i,s}=1] \times Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}+D_{i,s}=1])).\end{aligned}$$

Using the fact that $(Y_i(1,0), Y_i(0,1)) \perp\!\!\!\perp D_{i,t}, D_{i,s}|Z_{i,t}, Z_{i,s}$ (i.e. unconfoundedness conditional on unobservables), we have:

$$\begin{aligned}&\Delta(s) = \sum_{y,w\in\{0,1\}} \\ &(E[Y_i(1,0)|Z_{i,t}=y, Z_{i,s}=w] \times \\ &\;(Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}=1, D_{i,s}=0] - Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}+D_{i,s}=1]) - \\ &\;E[Y_i(0,1)|Z_{i,t}=y, Z_{i,s}=w] \times \\ &\;(Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}=0, D_{i,s}=1] - Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}+D_{i,s}=1])).\end{aligned}$$

To simplify this expression for the bias term first note that, applying Bayes' Rule, we can write for any (y, w):

$$\begin{aligned}&Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}=1, D_{i,s}=0] - Pr[Z_{i,t}=y, Z_{i,s}=w|D_{i,t}+D_{i,s}=1] \\ =& \frac{1}{Pr[D_{i,t}=0, D_{i,s}=1]Pr[D_{i,t}+D_{i,s}=1]} \times \\ &(Pr[D_{i,t}=1, D_{i,s}=0|Z_{i,t}=y, Z_{i,s}=w]Pr[D_{i,t}=0, D_{i,s}=1] - \\ &Pr[D_{i,t}=0, D_{i,s}=1|Z_{i,t}=y, Z_{i,s}=w]Pr[D_{i,t}=1, D_{i,s}=0]) \times Pr[Z_{i,t}=y, Z_{i,s}=w].\end{aligned}$$

Using Assumptions 1 and 2,

$$\begin{aligned}
&Pr[D_{i,t}=1, D_{i,s}=0|Z_{i,t}=Z_{i,s}=y] = Pr[D_{i,t}=0, D_{i,s}=1|Z_{i,t}=Z_{i,s}=y] \text{ for any } y\\
&Pr[D_{i,t}=1, D_{i,s}=0|Z_{i,t}=y, Z_{i,s}=w] = Pr[D_{i,t}=0, D_{i,s}=1|Z_{i,t}=w, Z_{i,s}=y] \text{ for } y \neq w\\
&Pr[D_{i,t}=1, D_{i,s}=0] = Pr[D_{i,t}=0, D_{i,s}=1].
\end{aligned}$$

It follows that

$$\begin{aligned}
&Pr[Z_{i,t}=Z_{i,s}=1|D_{i,t}=1, D_{i,s}=0] - Pr[Z_{i,t}=Z_{i,s}=1|D_{i,t}+D_{i,s}=1]\\
=&Pr[Z_{i,t}=Z_{i,s}=0|D_{i,t}=1, D_{i,s}=0] - Pr[Z_{i,t}=Z_{i,s}=0|D_{i,t}+D_{i,s}=1] = 0
\end{aligned}$$

and

$$\begin{aligned}
&Pr[Z_{i,t}=1, Z_{i,s}=0|D_{i,t}=1, D_{i,s}=0] - Pr[Z_{i,t}=1, Z_{i,s}=0|D_{i,t}+D_{i,s}=1]\\
=& -Pr[Z_{i,t}=0, Z_{i,s}=1|D_{i,t}=1, D_{i,s}=0] - Pr[Z_{i,t}=0, Z_{i,s}=1|D_{i,t}+D_{i,s}=1]\\
=& \frac{1}{Pr[D_{i,t}+D_{i,s}=1]} \times (p-q) \times \frac{f(s-t)}{2}.
\end{aligned}$$

Next note that a similar application of Bayes' Rule and Assumptions 1 and 2 establishes that

$$\begin{aligned}
&Pr[Z_{i,t}=Z_{i,s}=1|D_{i,t}+D_{i,s}=1] - Pr[Z_{i,t}=Z_{i,s}=1|D_{i,t}=0, D_{i,s}=1]\\
=&Pr[Z_{i,t}=Z_{i,s}=0|D_{i,t}+D_{i,s}=1] - Pr[Z_{i,t}=Z_{i,s}=0|D_{i,t}=0, D_{i,s}=1] = 0
\end{aligned}$$

and

$$\begin{aligned}
&Pr[Z_{i,t}=1, Z_{i,s}=0|D_{i,t}+D_{i,s}=1] - Pr[Z_{i,t}=1, Z_{i,s}=0|D_{i,t}=0, D_{i,s}=1]\\
=& -Pr[Z_{i,t}=0, Z_{i,s}=1|D_{i,t}+D_{i,s}=1] - Pr[Z_{i,t}=0, Z_{i,s}=1|D_{i,t}=0, D_{i,s}=1]\\
=& \frac{1}{Pr[D_{i,t}+D_{i,s}=1]} \times (p-q) \times \frac{f(s-t)}{2}.
\end{aligned}$$

Thus we have:

$$\begin{aligned}
\Delta(s-t) =& \frac{1}{Pr[D_{i,t}+D_{i,s}=1]} \times (p-q) \times \frac{f(s-t)}{2} \times\\
&(E[Y_i(1,0)+Y_i(0,1)|Z_{i,t}=1, Z_{i,s}=0] - E[Y_i(1,0)+Y_i(0,1)|Z_{i,t}=0, Z_{i,s}=1]).
\end{aligned}$$

Noting that $\lim_{s-t\to 0} f(s-t) = 0$ therefore completes the proof. □

Table 1— Institutional Details and Summary Statistics

	High Court		Court of Appeal (Civil Division)	
Institutional Feature				
Type of cases	Civil cases at first instance		Civil cases on appeal from High Court	
Number of designated judges[1]	108		37	
Size of panel hearing cases	1 judge		Typically 3 judges	
Decision taken by panel	Find in favour of plaintiff or respondent		Affirm or reverse first instance judgement	
Right of appeal[2]	Court of Appeal (Civil Division)		House of Lords	
Rank of judges who are *automatically* ticketed to hear cases	Justice and above		Lord Justice and above	
Rank of judges who can be *discretionally* ticketed to hear cases	Below Justice Retired Justice and above		Justice Retired Justice and above	
Criteria used to allocate cases to ticketed judges	Experience, legal specialism, availability		Cab-rank principle	
Duration of cases	Typically weeks		Typically 1 or 2 days	
Basic Summary Statistics				
Number of cases in full dataset	28307		15083	
Number of dissenting opinions	N/A		250	(1.7%)
Number of *linked* cases[3]	2262		2262	
No. of *linked* cases appealed[4]	2262		221	(9.7%)
No. of *linked* cases affirmed	1384	(61.2%)	111	(50.2%)
No. of *linked* cases reversed	878	(38.8%)	110	(49.8%)
No. of *linked* cases positively cited	194	(8.6%)	711	(31.4%)
No. of *linked* cases negatively cited	27	(1.2%)	122	(5.4%)

Notes:

1. Number of High Court judges (Justices) and Court of Appeal judges (Lord Justices) at the end of our sample in December 2005.
2. The High Court hears a small number of criminal cases on appeal from lower criminal courts. For these cases, the right of appeal lies directly to the House of Lords.
3. A case is classified as *linked* if: (i) Westlaw UK includes a link to the CA Civ (High Court) case in the Direct (Previous) History of the High Court (CA Civ) case, (ii) no relevant data fields are missing, and (iii) the CA Civ case takes place during term-time.
4. For the linked High Court cases, these are the linked CA Civ cases (reviews of the High Court judge's decision). For the linked CA Civ cases, these are subsequent reviews of the CA Civ judges' review decision in the House of Lords.

Table 2— Sample Size and Identification Concerns, by Length of Time Interval

Panel A.

	Total treatment level	Length of Time Interval 10 days (1) Obs.	 % full	 40 days (2) Obs.	 % full
Full sample	≥ 0	2262	100.0	2262	100.0
No treatment sample	= 0	2221	98.2	2159	95.4
Single treatment sample	= 1	34	1.5	54	2.4
Any treatment sample	≥ 1	41	1.8	103	4.6

Panel B.

	Length of Time Interval 10 days (1)	 40 days (2)
Single treatment sample		
Proportion treated in period Before	0.441	0.426
Proportion treated in period After	0.559	0.574
Difference in proportions	-0.118	-0.148
t-test for difference (*p*-value)	0.339	0.126
Any treatment sample		
Mean treatment level in period Before	0.634	0.913
Mean treatment level in period After	0.561	1.078
Difference in means	0.073	-0.165
t-test for difference (*p*-value)	0.611	0.299

Notes: Total treatment level counts the number of days in the (10 or 40-day) period immediately before, and the number of days in the (10 or 40-day) period immediately after, the date of the review on which there is a CA Civ judgement where the panel contains both the reviewed judge and one of his reviewers. The single treatment sample consists of observations where there is *exactly one* day, either in the (10 or 40-day) period immediately before or the (10 or 40-day) period immediately after the date of the review, where there is CA Civ judgement where the panel contains both the reviewed judge and one of his reviewers. The any treatment sample consists of observations where there is *at least one* day, either in the (10 or 40-day) period immediately before or the (10 or 40-day) period immediately after the date of the review, where there is CA Civ judgement where the panel contains both the reviewed judge and one of his reviewers.

Table 3— Balancing Tests, Length of Time Interval is 10 days

	Level Approach						Order Approach					
	Treated Before$^{10/0}$		Untreated Before$^{10/0}$		*t*-test *p*-value	Norm Diff	Treated Before$^{-10/0}$, Untreated After$^{0/10}$		Treated After$^{0/10}$, Untreated Before$^{-10/0}$		*t*-test *p*-value	Norm Diff
	(1)		(2)		(3)	(4)	(5)		(6)		(7)	(8)
Candidate Selection Variables	mean	(s.d)	mean	(s.d)			mean	(s.d)	mean	(s.d)		
Rank reviewed judge 1 day before decision:												
Below Justice	0	(0)	0.203	(0.403)	0.021	-0.503	0	(0)	0	(0)		0
Justice/retired	0.238	(0.436)	0.707	(0.455)	0.000	-0.744	0.333	(0.488)	0.421	(0.507)	0.614	-0.177
Above Justice	0.762	(0.436)	0.090	(0.286)	0.000	1.289	0.667	(0.488)	0.579	(0.507)	0.614	0.125
Rank reviewed judge 40 days after decision:												
Below Justice	0	(0)	0.203	(0.403)	0.021	-0.503	0	(0)	0	(0)		0
Justice/retired	0.238	(0.436)	0.704	(0.457)	0.000	-0.738	0.333	(0.488)	0.421	(0.507)	0.614	-0.125
Above Justice	0.762	(0.436)	0.093	(0.290)	0.000	1.278	0.667	(0.488)	0.579	(0.507)	0.614	0.125
Number of reviewers	3	(0)	2.832	(0.513)	0.134	0.327	3	(0)	3	(0)		0
Other Observables												
Time prior ruling to review (years)	0.868	(0.465)	0.939	(0.537)	0.546	-0.100	0.852	(0.482)	1.036	(0.492)	0.282	-0.267
Reviewer workload:$												
Total workload, Before$^{-10/0}$ + After$^{0/10}$	10.91	(4.230)	9.361	(4.769)	0.140	0.243	10.47	(4.274)	9.895	(2.865)	0.644	0.112
Workload before review, Before$^{-10/0}$	5.286	(2.667)	3.814	(2.867)	0.019	0.376	4.421	(2.341)	5.333	(2.870)	0.315	-0.246
Workload after review, After$^{0/10}$	3.571	(3.295)	3.607	(3.021)	0.957	-0.008	2.933	(2.712)	3.421	(2.652)	0.303	-0.129
Coverage of prior ruling:												
1[The Times Law Report]	0.714	(0.463)	0.336	(0.472)	0.000	0.572	0.800	(0.414)	0.421	(0.507)	0.026	0.579
1[The Independent Law Report]	0.095	(0.301)	0.120	(0.325)	0.728	-0.056	0.133	(0.352)	0.211	(0.419)	0.572	-0.143
Number of journal articles	3.667	(3.812)	3.037	(4.443)	0.518	0.108	3.667	(3.598)	3.263	(4.445)	0.777	0.071
Social ties with reviewed judge:												
1[At school together]	0	(0)	0.018	(0.134)	0.532	-0.134	0	(0)	0	(0)		0
1[At university together]	0.095	(0.301)	0.024	(0.152)	0.034	0.211	0	(0)	0	(0)		0
1[Same legal chambers]	0.190	(0.402)	0.058	(0.234)	0.010	0.610	0.200	(0.414)	0.105	(0.315)	0.454	0.183
1[Same social club]	0.238	(0.436)	0.033	(0.180)	0.000	0.435	0.200	(0.414)	0.053	(0.229)	0.196	0.311
Number of observations	21		2241		2262	2262	15		19		34	34

Notes: Norm Diff stands for normalised difference. This is equal to the difference in the mean of the covariate between the two groups divided by the square root of the sum of sample variance of the covariate in the two groups. The length of time interval is 10 days. Hence, Before$^{-10/0}$ (After$^{0/10}$) is a count of the number of days in the 10 day period (before respectively after) the review panel's decision, upon which there is a CA Civ judgement where the panel contains the reviewed judge and one of his reviewers.

$. This variable excludes interactions. It is a count of days in the specified period with a CA Civ judgement where the panel contains at least one reviewer but *not* the reviewed judge.

Table 4— Comparison of Means

		Affirmation	Difference	S.E.	*p*-value
		Rate	(1)-(2)	(1)-(2)	(1)=(2)
Panel A					
	Order Test				
(1)	Before$^{-10/0}$=1, After$^{0/10}$=0	0.533	-0.362	0.143	0.017
(2)	After$^{0/10}$=1, Before$^{-10/0}$=0	0.895			
Panel B					
	Backward-looking Level Test				
(1)	Before$^{-10/0}$=1, After$^{0/10}$=0	0.533	-0.019	0.155	0.901
(2)	Before$^{-10/0}$=0, After$^{0/10}$=0, OtherBefore$^{-10/0}$=1	0.553			
Panel C					
	Forward-looking Level Test				
(1)	After$^{0/10}$=1, Before$^{-10/0}$=0	0.895	0.346	0.129	0.010
(2)	After$^{0/10}$=0, Before$^{-10/0}$=0, OtherAfter$^{0/10}$=1	0.548			
Panel D					
	Backward-looking Placebo Test				
(1)	Before$^{-10/0}$=1, After$^{30/40}$=0	0.533	-0.012	0.206	0.954
(2)	After$^{30/40}$=1, Before$^{-10/0}$=0	0.545			
Panel E					
	Forward-looking Placebo Test				
(1)	After$^{0/10}$=1, After$^{30/40}$=0	0.857	0.274	0.152	0.082
(2)	After$^{30/40}$=1, After$^{0/10}$=0	0.583			

Notes: Before$^{10/0}$ and After$^{0/10}$ are defined as in Table 3. After$^{30/40}$ corresponds to the 10-day period staring 30 days after the review. OtherBefore$^{-10/0}$ (OtherAfter$^{0/10}$) is a count of the number of days in the 10 day period (before respectively after) the review panel's decision, upon which there is a CA Civ judgement where the panel contains the reviewed judge but *not* one of his reviewers.

Table 5— Robustness: Using Full Sample and Controlling for Observables

Y_i =1[Review panel *i* affirms prior ruling]	Linear Regression Models											
	Order Tests				Backward-Looking Level Test				Forward-Looking Level Test			
	(1)		(2)		(3)		(4)		(5)		(6)	
	Coeff	(s.e.)	Coeff	(s.e.)	Coeff	(s.e.)	Coeff	(s.e.)	Coeff	(s.e.)	Coeff	(s.e.)
Treatment level in:												
Before$^{-10/0}$	-0.334	(0.100) ***	-0.320	(0.111) ***	-0.077	(0.078)	-0.057	(0.087)				
After$^{0/10}$									0.239	(0.065) ***	0.236	0.070 ***
Before$^{-10/0}$ + After$^{0/10}$	0.246	(0.064) ***	0.245	(0.069) ***								
Proxy for Ticketing Status of Review Judge												
OtherBefore$^{-10/0}$							-0.022	(0.020)				
OtherAfter$^{0/10}$											0.019	(0.020)
Controls												
1[Reviewed judge below Justice at decision]			-0.078	(0.028) ***			-0.081	(0.028) ***			-0.078	(0.028) ***
1[Reviewed judge Justice/retired at decision]			Omitted				Omitted				Omitted	
1[Reviewed judge above Justice at decision]			-0.033	(0.038)			-0.006	(0.037)			-0.059	(0.043)
Number of reviewers			-0.051	(0.021) **			-0.048	(0.021) **			-0.052	(0.021) **
Time from prior ruling to review (years)			-0.010	(0.020)			-0.009	(0.020)			-0.011	(0.020)
Total reviewer workload, Before$^{-10/0}$ + After$^{0/10}$			0.002	(0.002)			0.002	(0.002)			0.002	(0.002)
1[The Times Law Report]			-0.016	(0.024)			-0.016	(0.023)			-0.017	(0.024)
1[The Independent Law Report]			-0.051	(0.033)			-0.050	(0.034)			-0.050	(0.034)
No. of journal articles			-0.002	(0.003)			-0.002	(0.003)			-0.002	(0.003)
1[Chancery]			0.105	(0.045) **			0.103	(0.045) **			0.106	(0.045) **
1[Civil]			0.069	(0.046)			0.066	(0.046)			0.070	(0.046)
1[Crime]			0.117	(0.072)			0.108	(0.072)			0.111	(0.073)
1[Employment]			Omitted				Omitted				Omitted	
1[Family]			0.132	(0.061) ***			0.137	(0.061) **			0.134	(0.060) **
1[Public]			0.147	(0.048) ***			0.150	(0.049) ***			0.149	(0.049) ***
1[At school together]			0.017	(0.077)			0.010	(0.077)			0.021	(0.076)
1[At university together]			0.039	(0.066)			0.042	(0.067)			0.034	(0.066)
1[Same legal chambers]			-0.023	(0.043)			-0.023	(0.043)			-0.025	(0.043)
1[Same social club]			-0.134	(0.057) *			-0.139	(0.057) **			-0.139	(0.057) **
Number of Observations	2262		2262		2262		2262		2262		2262	

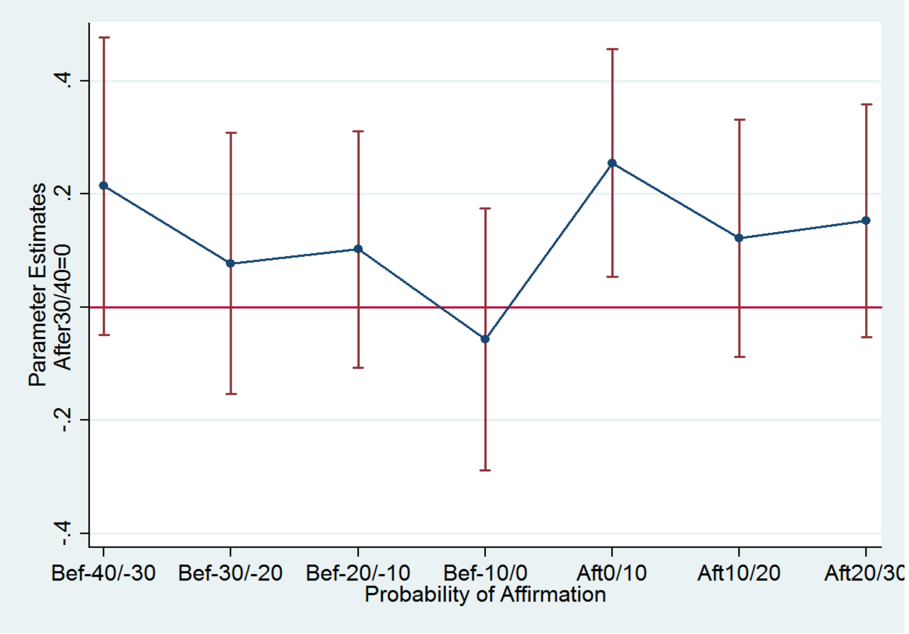
Notes: Robust standard errors in parentheses. ***, ** and * denote significance at 1, 5 and 10 percent levels respectively. The length of time unit is 10 days. Before$^{10/0}$ and After$^{0/10}$ are defined as in Table 3. OtherBefore$^{-10/0}$ and OtherAfter$^{0/10}$ are defined in Table 4.

Table 6—Forward-looking Placebo Test, By the Seniority of the Reviewer

		Affirmation Rate	Difference (1)-(2)	S.E. (1)-(2)	*p*-value (1)=(2)
Panel A					
	Reviewer senior to Reviewed Judge				
(1)	$\text{After}^{0/10}=1$, $\text{After}^{30/40}=0$	0.667	-0.083	0.243	0.734
(2)	$\text{After}^{30/40}=1$, $\text{After}^{0/10}=0$	0.750			
Panel B					
	Reviewer same rank as Reviewed Judge				
(1)	$\text{After}^{0/10}=1$, $\text{After}^{30/40}=0$	1.000	0.500	0.185	0.011
(2)	$\text{After}^{30/40}=1$, $\text{After}^{0/10}=0$	0.500			

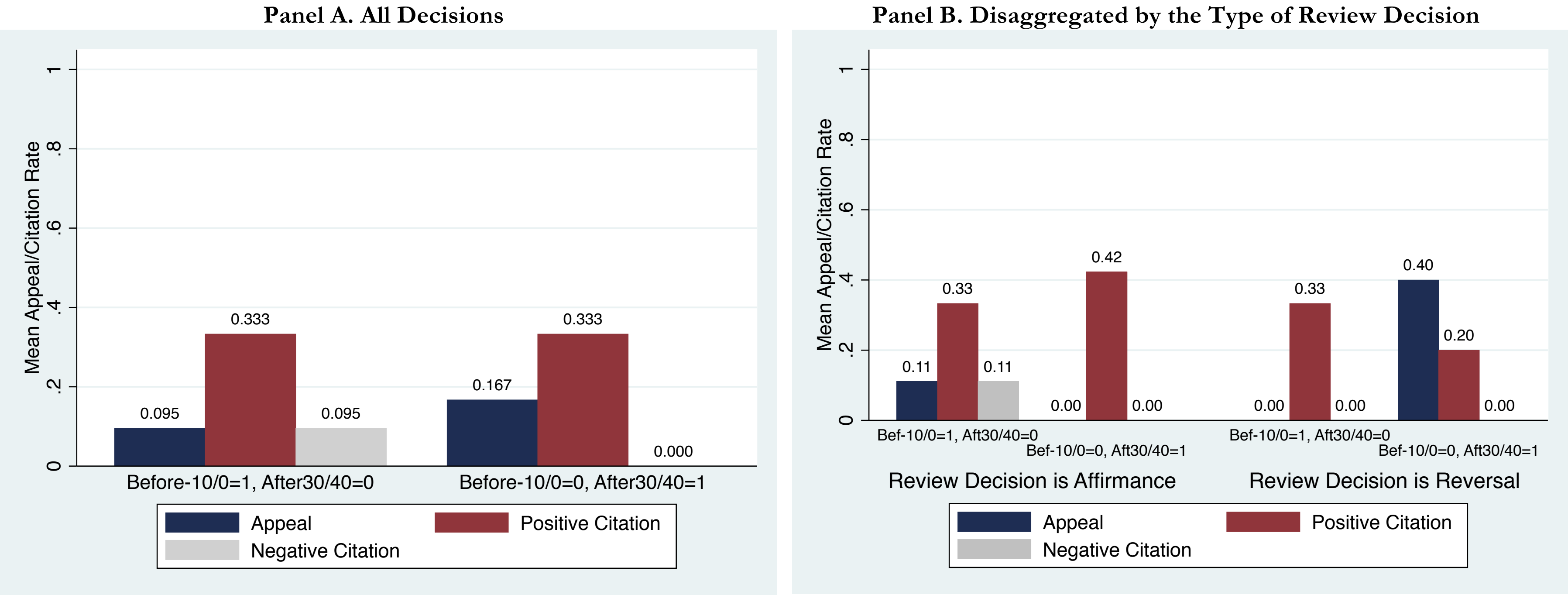
Notes: $\text{After}^{0/10}$ and $\text{After}^{30/40}$ are as defined in Table 4. We split the sample in Table 4 Panel C into two sub-samples, depending on the relative seniority (at the time of the review) of the reviewer and reviewed judge that experience the on-the-job interaction. There are no observations where the reviewer with the on-the-job interaction is less senior than the reviewed judge.

Figure 1— Predicted Probability of Affirmation by Time of Treatment



Notes: The figure depicts the point estimates from (8). Point estimates and confidence intervals are plotted. The horizontal line at zero illustrates the fact that all the estimates are relative to the After$^{30/40}$ group.

Figure 2— The Quality of Review Decisions



Panel A. All Decisions

Panel B. Disaggregated by the Type of Review Decision

Notes: The unit of time is set at 10 days. In Panel A, the bars labelled 'Before$^{-10/0}$=1, After$^{30/40}$=0' depict the mean appeal (or citation) rate for observations where there is one day in the 10 days starting immediately after the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but no day in the 10 days *starting 30 days after* the review where the panel contains the reviewed judge and one of his reviewers. The bars labelled 'Before$^{-10/0}$=0, After$^{30/40}$=1'depict the mean appeal (or citation) rate for observations where there is no treatment in the 10 days starting immediately after the review but a single treatment in the 10 days *starting 30 days after* the review. Panel B disaggregates by the type of review decision (i.e. an affirmance or reversal). For the observations where the review decision is a reversal, the difference in mean appeal rate between the treated and placebo groups of 0.400 is significant at the 10 percent level (*p*=0.07).