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

Jordi Blanes i Vidal† Clare Leaver‡

April 10, 2013

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 leniency 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 and Michael Whinston. 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, favouritism motivated by personal ties, leniency driven by a fear of awkwardness and/or reprisal, or all three. As we discuss below, there have been few attempts to investigate this issue empirically. To the best of our knowledge, only a small number of randomised controlled 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 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 leniency 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 professional settings, such

[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".

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

[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.

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 $t+s$, 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), 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 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 favouritism motivated by personal ties (respectively leniency driven by a fear of awkwardness and/or reprisal).[6]

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 leniency mechanism in greater detail. First, we substantiate the rationale for leniency by providing direct 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 leniency bias when assessing junior colleagues than when assessing peers of the same rank. Finally, we draw out additional empirical predictions relating to the

[6]Strictly speaking, a negative difference is also consistent with the predominant force being antipathy or overwork caused by an interaction before the review. We use additional 'placebo' tests based on unanticipated future interactions, as well as the workload of the reviewers/reviewed judge, to distinguish leniency from these competing mechanisms.

quality of review decisions. Developing a simple theoretical framework, we show how an anticipated on-the-job interaction can cause a leniency bias in decision-making 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 *a priori* claims (e.g. Robertson 1976), open peer-review can be subject to leniency 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 leniency bias when assessing junior colleagues suggests that firms should reconsider the merits of *decentralised* open performance appraisals, and highlights the need for anonymity in '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.[7] 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.

[7]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.

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.[8]

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.[9]

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.[10] In contrast, reviewed judges who held the post of Lord Justice at the time of their judgement will be automatically

[8]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.

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

[10]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.[11] 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. 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 leniency 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.

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

[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 $t \neq 0$ and $s > 0$,

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

This assumption states that, if the realisation of the selection variable is the same at two dates t and $t + s$, 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

[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.

because 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 $t \neq 0$ and $s > 0$,

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

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

This assumption states that the likelihood of the selection variable taking a different value at two dates t and $t + s$ is increasing in the elapsed time s 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 $t + s$. 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)

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

[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 $t+s$ (another level effect)

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

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

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

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

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

we can state the following result.

Proposition 1. *Under Assumptions 1 and 2,*

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

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 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 $E[Y_i(D_{i,t}=1, D_{i,t+s}=0)]$ from conditioning on $D_{i,t}=1, D_{i,t+s}=0$ rather than $D_{i,t}+D_{i,t+s}=1$ is greater than the increase in $E[Y_i(D_{i,t}=0, D_{i,t+s}=1)]$ from conditioning on $D_{i,t}=0, D_{i,t+s}=1$ rather than $D_{i,t}+D_{i,t+s}=1$. 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 $Z_{i,t} = Z_{i,t+s} = 1$ or the likelihood that $Z_{i,t} = Z_{i,t+s} = 0$. True, conditioning on $D_{i,t}=1, D_{i,t+s}=0$ rather than $D_{i,t}+D_{i,t+s}=1$ increases the likelihood that $Z_{i,t}=1, Z_{i,t+s}=0$ and decreases the likelihood that $Z_{i,t}=0, Z_{i,t+s}=1$. However, as s becomes small, the likelihood that $Z_{i,t} \neq Z_{i,t+s}$ (and hence the magnitude of any bias) vanishes. Thus, as s becomes small, unobservables will almost certainly stay fixed at either $Z_{i,t} = Z_{i,t+s} = 1$ or $Z_{i,t} = Z_{i,t+s} = 0$ 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 Favouritism and Leniency

To illustrate how Proposition 1 can be used to explore the hypotheses discussed in the Introduction, suppose that $t = -1$ and $s = 2$. If the appeals process is influenced by on-the-job interaction and the mechanism is *favouritism* (judicial sympathy), then for at least one i the unit causal effect in (1) should be positive.[18] That is, there should be at least one panel where the potential outcome that would be realised given an interaction on the date before (but not after) the review is an affirmance but the potential outcome that would be realised given no interaction on either date is a reversal. However, if the appeals process is open to influence by on-the-job interaction and the mechanism is *leniency*, then for at least one i the unit causal effect in (2) should be positive. Hence, if favouritism is the predominant force, then the difference between these effects (the unit causal effect in (3)) should be positive. Alternatively, if leniency is the predominant force, then this effect of treatment order should be negative.[19]

Now consider a comparison of the mean realised outcome for units treated with an interaction on the date before (but not after) the review with the mean realised outcome for units treated with an interaction on the date after (but not before) the review. Applying Proposition 1, this difference is approximately equal to $\tau_{-1,2}$, the average of the unit causal effect in (3) for the subpopulation treated once. If this comparison of means is significantly different from zero, then the unit causal effect in (3) must be non-zero for at least one i. Such a finding would be sufficient to reject the hypothesis that the appeals process is above the influence of on-the-job interaction. A positive estimate would be consistent with favouritism being the predominant force, while a negative estimate would be consistent with leniency being the predominant force. The tests summarised in Section 4 below are based on these observations, although, of course, we allow for the possibility of observable selection variables, as well as the possibility that treatment can occur on more than two dates.

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

[18] The *favouritism* mechanism relies on friendship among co-workers becoming established relatively quickly, i.e. through a single job interaction. If this is not the case, then we will not find a positive effect in our empirical analysis.

[19] This effect of treatment order would also be negative if on-the-job interaction at $t = -1$ engendered *antipathy* rather than sympathy. In Section 4.2, we explain how 'placebo' treatments can be used to test separately for the presence of antipathy and for the presence of leniency. The results of these placebo tests are reported in Section 5 and enable us to adjudicate between competing mechanisms.

3.4 Assessing Unconfoundedness

Our empirical strategy rests on the claim that, under Assumption 1 and 2 and for sufficiently small s, 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, 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.

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.[20]

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 between $t = -1$ and $t = 1$ 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,

[20] 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.

we are forced to expand the time index for our treatment indicators beyond a single day.

We proceed by constructing three samples: one where the length of the time unit t is set at 10 days, another where it is set at 20 days, and a third where it is set at 40 days. Table 2 Panel A illustrates the associated sample sizes, along with a sample where t is set at 5 days for comparison purposes. Column 2 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 t is expanded as Column 3 and 4 confirm. In the following two subsections, we illustrate our estimation methods for the case where the length of t is set at 10 days.

4.2 Comparison of Means

Let $S_{i,-1}$ 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 and one of his reviewers, and $S_{i,1}$ 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 and one of his reviewers. Assuming that we have a random sample on Y_i, $S_{i,-1}$ and $S_{i,1}$ from the population with $S_{i,-1} + S_{i,-1} = 1$, the treatment effect $\tau_{-1,2}$ can be consistently estimated by a simple regression of the realised outcome Y_i on a constant and $S_{i,-1}$.[21] The results of this exercise,

$$\widehat{\tau}_{-1,2} = E[Y_i|S_{i,-1} = 1, S_{i,1} = 0] - E[Y_i|S_{i,-1} = 0, S_{i,1} = 1], \tag{4}$$

can be obtained simply by observing the raw data which we display in Figure 1.

In Section 2 we noted that, at the time of the review decision, panel members are unlikely to know who they will be working with at hearings taking place more than a month in the future. This observation suggests the following 'placebo' tests. Let $S_{i,4}$ denote the number of days in the 10 day period starting 30 days after the review and suppose we obtain the following estimates:

$$\widehat{\tau}_{-1,5} = E[Y_i|S_{i,-1} = 1, S_{i,4} = 0] - E[Y_i|S_{i,-1} = 0, S_{i,4} = 1], \tag{5}$$

and

$$\widehat{\tau}_{1,3} = E[Y_i|S_{i,1} = 1, S_{i,4} = 0] - E[Y_i|S_{i,1} = 0, S_{i,4} = 1]. \tag{6}$$

Applying the logic from Section 3.3, if the appeals process is influenced by on-the-job interaction and

[21] Note that, in the single treatment sample, the count variables coincide with the binary treatment indicators and (hence) $S_{i,-1} = 1$ implies $S_{i,1} = 0$ and $S_{i,-1} = 0$ implies $S_{i,1} = 1$.

the force is favouritism, then $\widehat{\tau}_{-1,2}$ and $\widehat{\tau}_{-1,5}$ should be of a similar positive magnitude. This is because, under the favouritism hypothesis, future interactions have no effect, irrespective of whether they can be anticipated at the time that the panel takes its decision. On the other hand, if the force is leniency, then $\widehat{\tau}_{-1,2}$ and $\widehat{\tau}_{1,3}$ should be of a similar absolute magnitude but the former should be negative while the latter should be positive. This is because, under the leniency hypothesis, past interactions and *unanticipated* future interactions have no effect. The results of this exercise can be obtained simply by observing the raw data which we display in Figure 2.

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:

$$Y_i = \alpha + \beta \cdot S_{i,-1} + \gamma \cdot (S_{i,-1} + S_{i,1}) + \zeta' \mathbf{X}_i + \varepsilon_i \tag{7}$$

for $i = 1, ..., 2262$. The model in (7) imposes two additional assumptions, namely that the effect of treatment is linear and (more restrictively) is constant across i. Under these assumptions, $\beta = \tau_{-1,2}$ and so the OLS estimate $\widehat{\beta}$ provides a robustness check for (4). The results of this estimation exercise are presented in Table 4. To control for the possibility of treatment in other periods, we also estimate:

$$Y_i = \theta + \sum_{t=-4}^{3} \beta_t \cdot S_{i,t} + \phi \cdot \sum_{t=-4}^{4} S_{i,t} + \xi' \mathbf{X}_i + \epsilon_i \tag{8}$$

for $i = 1, ..., 2262$. Since $S_{i,4}$ is omitted, $\widehat{\beta}_{-1}$ and $\widehat{\beta}_1$ are essentially robustness checks for (5) and (6). The results of this exercise are presented in Table 5 and Figure 3.

5 Results

We now summarise our results, postponing any interpretation until Section 5.4.

5.1 Assessing Unconfoundness

Table 2 Panel B reports the tests for equal proportions of orders of treatment using the single treatment samples. Column 2 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. Column 3 shows that the proportion of panels treated once in the 20 days before (but not after) the review is identical to the proportion treated once in the 20 days after (but not before) the review. Column 4 shows that the proportion of panels treated once in the 40 days before (but not after) the review is lower than the proportion treated once in the 40 days after (but not before) the review but, again, a $t-$test fails to reject the null hypothesis of equal proportions. 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, -0.074 and -0.165 (for the 10, 20 and 40 time lengths respectively) but, again, in every case $t-$tests fail to reject the null hypothesis of equality.

Table 3 reports the results of balancing tests where, to be conservative, we set the length of t to 40 days. 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 $S_{i,-1}$. As expected, there is a large and statistically significant difference in the rank of the reviewed judge across the treated ($S_{i,-1} > 0$) and untreated ($S_{i,-1} = 0$) groups. For instance, 74 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 8 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 $S_{i,-1}$ and $S_{i,1}$. The difference in the rank of the reviewed judge across the treated before but not after group ($S_{i,-1} = 1, S_{i,-1} = 0$) and the treated after but not before group ($S_{i,-1} = 0, S_{i,-1} = 1$) is now far smaller. For instance, 65 percent of the group treated at time -1 but not at time 1 review a judge who holds a post above the rank of justice at the decision date, while the corresponding figure for the group treated at time 1 but not at time -1 is 58 percent. The difference of 7 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 (Imbens and Rubin, forthcoming). Reassuringly, the normalised difference drops substantially, from 1.266 in column 4 to 0.109 in column 8. The difference in the average number of reviewers remains small and is close to 3 for both groups.[22] Only the coverage of the first instance judgement in newspaper law reports appears to 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.[23]

5.2 Comparison of Means

Time unit set at 10 days The first bar in Figure 1 Panel A 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 (i.e. at $t = -1$ but not at $t = 1$) is 0.533. In stark contrast, the second bar shows that the mean affirmance rate for the group treated once in the 10 days after, but not in the 10 days before, the review (at $t = 1$ but not at $t = -1$) is 0.895. The first plot in Panel B confirms that the estimated difference in means, $\widehat{\tau}_{-1,2} = -0.362$, is statistically significant at 10 percent (this is also true at 5 percent).

Figure 2 reproduces the results from Figure 1, together with three 'placebo' tests. The first test is a simple robustness check for the order approach used in Figure 1. Here, we compare the affirmance rate for a group where the reviewed judge experiences an on-the-job interaction with other CA civ judges (not her reviewers) shortly before the review with the affirmance rate for a group where the reviewed judge experiences an on-the-job interaction with other CA Civ judges (not her reviewers) shortly after the review. Since the reviewers do not interact with the reviewed judge, there should be no favouritism or leniency effect. Moreover, since we are testing for an order rather than level effect of on-the-job interaction, there should also be no selection bias. Rejection of the null hypothesis of equal means would therefore cast doubt on our order approach. The third bar in Figure 2 Panel A shows that the mean affirmance rate for the group with no treatment in the 10 days before or the 10 days after the review but where the reviewed judge works in a CA Civ panel in the 10 days before (but not after) the review is 0.552. Reassuringly, the fourth bar shows that the mean affirmance rate for the group with no treatment in the 10 days before or the 10 days after the review but where the reviewed judge works in a CA Civ panel in the 10 days after (but not before) the review is similar at 0.548. The second plot in Panel B shows that the estimated difference in means, 0.004, is not statistically different from zero at standard inference levels.

[22]The fact that the normalised difference actually rises a little moving from column 4 to 8 is surprising but, with the estimate below the rule of thumb of a quarter suggested by Imbens and Rubin, this is not particularly worrying.

[23]Table A1 confirms that the balancing test results are qualitatively similar when we set the length of t to 10 days.

The purpose of the remaining placebo tests is to shed light on the relative importance of the favouritism and leniency mechanisms. Following the discussion in Section 4.2, we do this by comparing the affirmance rate for groups that receive a genuine treatment (an on-the-job interaction between a reviewer and the reviewed judge shortly before/after the review) with the affirmance rate for groups that receive a placebo treatment (an on-the-job interaction between a reviewer and the reviewed judge a month after the review). The fifth bar in Panel A shows that the mean affirmance rate for the group treated once in the 10 days before the review but not in the 10 days *starting 30 days after* the review is 0.533. The sixth bar shows that the mean affirmance rate for the group treated once (with a placebo) in the 10 days starting 30 days after the review but not in the 10 days before the review is practically identical at 0.545. The third plot in Panel B confirms that the estimated difference in means, $\widehat{\tau}_{-1,5} = -0.012$, is not significantly different from zero at standard inference levels.

Turning to the third test, the seventh bar in Figure 2 Panel A shows that the mean affirmance rate for the group treated once in the 10 days immediately after the review but not in the 10 days starting 30 days after the review is 0.857. In contrast, the final bar shows that the mean affirmance rate for the group treated (with a placebo) in the 10 days starting 30 days after the review but not in the 10 days starting immediately after the review is 0.583. The final plot in Panel B confirms that this estimated difference in means, $\widehat{\tau}_{1,3} = 0.274$, is statistically significant at 10 percent and, notably, is similar in absolute magnitude to the estimate $\widehat{\tau}_{-1,2} = -0.362$.

Time unit set at 20 and 40 days The third bar in Figure 1 Panel A shows that the mean affirmance rate for the group treated once in the 20 days before, but not in the 20 days after, the review is 0.571. The fourth bar shows that the mean for the group treated once in the 20 days after, but not in the 20 days before, the review is 0.761. As the second plot in Figure 1 Panel B illustrates, the estimated difference in means, -0.190, is economically significant but just shy of significance at 10 percent. The fifth and sixth bars repeat this exercise for the 40 days before and after the review. In this case, the estimated difference in means is not economically or statistically significant.

5.3 Robustness Checks using Regression Models

Time unit set at 10 Days The first two columns in Table 4 report estimates of the marginal 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, -0.362, in Figure 1 Panel B. Again, a $t-$test rejects the null of a zero effect at 1 percent. Column 2 is based on

the specification in (7) and controls for the candidate selection variables, as well as other observable characteristics. The estimated marginal 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 affirmation.

The first two columns in Table 5 are based on the specification in (8) and report estimates of the marginal effect of an additional treatment taking place at a time $t = -4, ... -1, 1, ..., 3$ rather than at $t = 4$. The coefficients in column 1 are marginal effects from a linear specification, while the coefficients in column 2 are odds ratios from a logistic specification. Of the treatment coefficients, only $\widehat{\beta}_1$ is significantly different from zero, indicating that receipt of an additional treatment *rather than a placebo* is associated with an increase in the probability of affirmance only if it occurs in the 10 days immediately after a review.

To facilitate visual comparisons, we also compute predicted probabilities of affirmance. The first bar in Figure 3 Panel A plots the predicted probability of affirmance when $S_{i,-4} = 1$ and $S_{i,t} = 0$ for all $t = -3, ..., 4$, with all other variables held at the mean for a panel with a single treatment. The remaining bars do likewise but with the single treatment taking place in the stated time period. The predicted probability of affirmance at $t = 1$ barely changes, dropping from 0.857 in Figure 2 Panel A to 0.815 in Figure 3 Panel A, confirming that controlling for treatment in other periods also does little to change our main result.

It is evident from Figure 3 Panel A that the predicted probability of affirmance is also high at time $t = -4$ (the 10 day period starting 40 days before the review). Although the effect of receiving an additional treatment in this period rather than a placebo is not significantly different from zero, it is possible that there could be significant treatment effects further back in time. The third column in Table 5 explores this possibility and reports estimates of an additional treatment taking place at time $t = -10, ... -1, 1, ..., 3$ rather than at $t = 4, ..., 10$. Reassuringly, none of the additional treatment coefficients are significant. The predicted probabilities of affirmance are displayed in Figure 3 Panel B.

Time unit set at 20 and 40 days The middle two columns in Table 4 report estimates of the marginal effect of an additional treatment taking place in the 20 days before the review rather than in the 20 days after the review. The estimated marginal effect in column 3 (without controls), -0.151, is a little lower than the estimated treatment effect -0.190 in Figure 1 Panel B but, given the larger sample size, is now significant at 5 percent. Controlling for observables again has little impact, with the estimated marginal effect dropping only slightly in column 4. The final two columns in Table 4 report

estimates of the marginal effect of an additional treatment taking place in the 40 days before the review rather than in the 40 days after the review. Consistent with the comparison of means, the estimated coefficients with and without controls are insignificant at standard inference levels.

Comparison of Order and Level Effects We conclude this subsection by comparing our results with a naive 'level' approach. Column 1 in Panel A of Table A2 repeats column 1 in Table 4 where the coefficient of interest is an estimate of the marginal effect of an additional treatment taking place at $t = -1$ rather than at $t = 1$ (an order effect). Column 2 adds a dummy for the rank of the reviewed judge at the time of the review decision. Despite the fact that the marginal effect of higher rank is positive and strongly significant, the coefficient of interest is unchanged at two decimal places. In columns 3 and 4 the coefficient of interest is an estimate of the marginal effect of an additional treatment taking place at $t = -1$ (a level effect), while in columns 5 and 6 the coefficient of interest is an estimate of the marginal effect of an additional treatment taking place at $t = 1$ (another level effect). For both sets of regressions, controlling for the rank of the reviewed judge *decreases* the magnitude of the estimated level effect. These findings lend further support to our claim that unobservables related to the quality of the reviewed judge are likely to create a positive omitted variable bias under a 'level' approach but not under the 'order' approach set out in Section 4. Of course, the magnitude of the order effect may still exceed each level effect if, as appears to be the case here, the level effects have opposing signs.

The key insight in our identification strategy is that, under Assumptions 1 and 2, conditioning on one order of treatment rather than another (say $D_{i,t} = 1, D_{i,t+s} = 0$ rather than $D_{i,t} = 0, D_{i,t+s} = 1$) has no impact on the likelihood that selection variables are fixed at one level rather than another (say at $Z_{i,t} = Z_{i,t} = 1$ rather than $Z_{i,t} = Z_{i,t} = 0$) and so these selection variables can safely be ignored when estimating the effect of treatment order (providing s is small). Although we believe that this strategy is the most credible way to investigate the favouritism and leniency hypotheses, for robustness we also compare this approach with an alternative strategy where we attempt to estimate the effect of treatment level by controlling for selection variables −notably ticketing status− directly.

Although historical lists of ticketed judges are not available, it is possible to construct a rough proxy using our data on panel composition. Specifically, if we observe a CA Civ judgement where the panel contains the reviewed judge at time t, this tells us that this judge must be ticketed at time t.[24] The second and third columns in Panel B of Table A2 repeat the analysis in Panel A including this (imperfect) control for ticketing status. The coefficients of interest are estimates of the marginal effect

[24]This is a rough proxy because absence of a CA Civ judgement where the panel contains the reviewed judge at time t does not imply that this judge is not ticketed at time t.

of an additional treatment taking place at $t = -1$ (columns 3 and 4) and at $t = 1$ (columns 5 and 6). In both columns, the coefficient on the proxy for ticketing status is insignificant, suggesting that it may not adequately capture selection. Despite this, the coefficients of interest are now stable to the inclusion of the rank of the reviewed judge. In particular, the negative coefficient on $S_{i,-1}$ remains insignificant, while the positive coefficient on $S_{i,1}$ remains strongly significant and unchanged at two decimal places. While it is unlikely that these are reliable estimates of the true level effects, the signs and magnitudes of the coefficients do offer reassurance that the 'order' effect reported in column 2 is not unduly large.

5.4 Interpretation

Time unit set at 10 Days We interpret the finding that $\widehat{\tau}_{-1,2}$ is significantly different from zero as evidence that (3) is non-zero for at least one i and hence 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 indicate 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 regression models, controlling for observables has little effect. Moreover, if there is selection bias, then it should increase with the elapsed time s but, rather than finding that $\widehat{\tau}_{-1,5}$ is bigger in absolute magnitude than $\widehat{\tau}_{-1,2}$, we find the reverse.

The fact that $\widehat{\tau}_{-1,2}$ is negative is consistent with the predominant force being *leniency* motivated by a fear of awkwardness and/or reprisal in the future interaction. An alternative interpretation is that past interaction engenders antipathy, rather than favouritism, towards the reviewed judge. This interpretation is consistent with our findings in Figure 1 and Table 4. However, the evidence from the second and third 'placebo' tests in Figure 2 suggests that such antipathy is unlikely. Under this alternative antipathy interpretation, $\widehat{\tau}_{-1,5}$ should be negative. This is because the antipathy explanation predicts the affirmation rate to be lower for the group treated just before the review than for the group treated with an unanticipated future interaction.[25] For similar reasons, $\widehat{\tau}_{1,3}$ should be zero. In contrast to these predictions, however, we are unable to reject the null hypothesis that $\widehat{\tau}_{-1,5}$ is zero, while we find that $\widehat{\tau}_{1,3}$ is positive. In fact, consistent with the leniency interpretation, $\widehat{\tau}_{1,3}$ is very similar in absolute magnitude to $\widehat{\tau}_{-1,2}$.[26]

[25] Since antipathy affects $E[Y_i|S_{i,-1} = 1, S_{i,1} = 0]$, this prediction is valid regardless of whether antipathy has a permanent or a short-term effect.

[26] Another interpretation is that a panel's recent workload (rather than its recent interaction with the reviewed judge) makes it more reversal prone, perhaps due to exhaustion or over-confidence. Figure A1 suggests that this alternative

Time unit set at 20 and 40 Days Repeating our baseline comparison of means (based on (4)) and regression analysis (based on (7)) with the length of the time unit set at 20 days yields qualitatively identical, but quantitatively smaller, results. As Table 4 indicates, doubling the length of t from 10 to 20 days roughly halves the estimated marginal effect. Guided by Figure 3, we attribute this to discounting. That is, on-the-job interaction has less of an effect when the panel expects it to occur further in advance of the review. Consistent with this interpretation, when we extend the length of t to 40 days, the estimated treatment and marginal effects are not significantly different from zero.

6 Exploring the Leniency Mechanism

Having argued that leniency could explain our results, we now assess this mechanism in more detail.

6.1 The Rationale for Leniency

We begin by looking for evidence to substantiate the rationale for leniency, namely that panel members will anticipate that a lenient decision (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.[27]

The results of our study of working relationships during on-the-job interactions are presented in Figure 4. Since dissents are rare events, we expand the sample size by setting the unit of time to 40 days. Panel A pools across all review decisions. The first bar shows 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

explanation is unlikely. All four bars in Panel A depict observations with the same workload, namely one CA Civ judgement where the panel contains a reviewer and that takes place either in the 10 days before the review (bars one and three) or in the 10 days immediately after the review (bars two and four). In the first two bars, the reviewed judge is present, while in the last two bars the reviewed judge is absent. Panel B shows that there is *not* a significant difference in mean affirmance rates when the reviewed judge is absent. We have also re-run the second and third placebo tests in Figure 2 using on-the-job interactions between the reviewed judge and *other* CA Civ judges (not her reviewers) as the placebo. Again, we find no significant effect of treatment before the review and a positive effect of treatment after the review, providing further evidence that it is leniency rather than antipathy that is driving our results.

[27] Table 1 shows that a dissenting opinion features in less than 2 percent of the 15083 CA Civ cases in our database.

in the on-the-job interaction is 7 percent. The second bar shows that, 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. Panel B disaggregates by the type of review decision. The first two bars indicate that the dissent rate is lower, by six percentage points, when the on-the-job-interaction occurs after rather than before an affirmation. This is consistent with a rationale for leniency. 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). Of course, power is an obvious concern here, particularly for the latter comparison. As such, we view these results as suggestive of a rationale for leniency.

6.2 Heterogeneity in Leniency

If leniency really is the mechanism at work in Figure 1 and Table 4, 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 leniency 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 leniency, 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 A1 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) at this judge. We can, however, look for heterogeneity along the seniority dimension. Since our objective is to explore the effects of leniency, 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 therefore split the observations into 'more senior' and 'same rank' subsamples.

Our results are presented in Figure 5. The first bar in 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 second bar shows that 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. As Panel B illustrates, the difference in means, -0.083, is both economically and statistically insignificant, indicating that we have failed to find evidence of a leniency effect in this 'more senior' subsample. The third bar in 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 holds the same rank as himself is 1.000. The final bar shows that 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 leniency 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 Leniency

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).[28] 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:

$$p_H > \frac{(C+D)\mu}{C(2\mu-1)+D} > p_L > \mu. \tag{9}$$

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: $-(1-\gamma_{p,s,r}) \cdot D - 1\left[r = 1, \sigma = 1\right] \cdot C$.

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 *leniency 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 re-arranging 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.

[28] 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.

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

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

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

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

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

There is a leniency 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.[29]

Our results are presented in Figure 6. Since our objective is to explore the effects of leniency, we again focus on groups that are treated with a single interaction *after* the review. 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 10 days starting 30 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 10 days starting 30 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 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 6 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,

[29]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.

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 appeal 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 tentative evidence to support Proposition 2 Parts ii and iii, and stronger evidence to support with Proposition 2 Part iv. These findings further substantiate our claim that anticipated on-the-job interaction can introduce a leniency 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 favouritism, leniency, 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 leniency 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 ($\widehat{\tau}_{-1,2} = -0.362$), or after the review but is unanticipated ($\widehat{\tau}_{1,3} = 0.274$). 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 leniency mechanism further, we present a model of leniency bias that yields additional 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.

Leniency bias seems unlikely 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 them disappears. 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.[30] Naturally, the benefit of reduced leniency 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).[31]

[30]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 leniency bias when the reviewed judge holds the rank of Justice at the time of the review this further step may not be necessary.

[31]Caution would be especially warranted if it is difficult to discount any leniency 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.

Turning to the generalisability and wider policy implications 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 can, in principle, work alongside their reviewers following a promotion 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 leniency bias in other professional settings. The leniency 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). Figure 5 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. 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.[32]

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

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.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

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

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

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

[12] 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.

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

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

[15] 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.

[16] Griffith, John A. G. 'The Politics of the Judiciary', 1997, London: Fontana Press.

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

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

[19] The Judges' Council 'Response to the Consultation Papers on Constitutional Reform', 2003, mimeo.

[20] 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.

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

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

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

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

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

[26] 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.

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

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

[29] 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.

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

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

[32] 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.

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

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

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

[36] 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.

[37] 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.

[38] 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 $\tau_{t,s}$, the bias term, $\Delta(s)$, is:

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

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

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

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

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

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

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

Using Assumptions 1 and 2,

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

It follows that

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

and

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

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

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

and

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

Thus we have:

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

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

Proof of Proposition 2. First, recall the panel's decision-making behaviour. If (i) $p = p_H$, $s = \sigma = 1$, (ii) $p = p_H$, $s = 1$, $\sigma = 0$, or (iii) $p = p_L$, $s = 1$, $\sigma = 0$, then the panel chooses $r = 1$. Otherwise, the panel chooses $r = 0$. Given this behaviour, it is straightforward to establish the following probabilities.

Part i. The probability of the review decision being an affirmation conditional on σ is:

$$\begin{aligned}\Pr[r=0|\sigma=1] &= \Pr[p=p_L] + \Pr[p=p_H] \times [\mu p_H + (1-\mu)(1-p_H)] \\ \Pr[r=0|\sigma=0] &= \Pr[p=p_L] \times [\mu p_L + (1-\mu)(1-p_L)] + \Pr[p=p_H] \times [\mu p_H + (1-\mu)(1-p_H)].\end{aligned}$$

Thus

$$\Pr[r=0|\sigma=1] - \Pr[r=0|\sigma=0] = \Pr[p=p_L] \times [\mu(1-p_L) + (1-\mu)p_L] > 0.$$

Part ii. The probability of the review decision being incorrect conditional on σ is:

$$\begin{aligned}\Pr[r\neq x|\sigma=1] &= \Pr[p=p_L] \times (1-\mu) + \Pr[p=p_H] \times [\mu(1-p_H) + (1-\mu)(1-p_H)] \\ \Pr[r\neq x|\sigma=0] &= \Pr[p=p_L] \times [\mu(1-p_L) + (1-\mu)(1-p_L)] \\ &\quad + \Pr[p=p_H] \times [\mu(1-p_H) + (1-\mu)(1-p_H)].\end{aligned}$$

Thus

$$\Pr[r\neq x|\sigma=1] - \Pr[r\neq x|\sigma=0] = \Pr[p=p_L] \times (p_L - \mu) > 0.$$

Part iii. The probability of the review decision being incorrect conditional on $r=0$ and σ is:

$$\begin{aligned}&\Pr[x=1|r=0,\sigma=1] = \frac{\Pr[p=p_L] \times (1-\mu) + \Pr[p=p_H] \times [(1-\mu)(1-p_H)]}{\Pr[p=p_L] + \Pr[p=p_H] \times [\mu p_H + (1-\mu)(1-p_H)]} \\ &\Pr[x=1|r=0,\sigma=0] \\ &= \frac{\Pr[p=p_L] \times [(1-\mu)(1-p_L)] + \Pr[p=p_H] \times [(1-\mu)(1-p_H)]}{\Pr[p=p_L] \times [\mu p_L + (1-\mu)(1-p_L)] + \Pr[p=p_H] \times [\mu p_H + (1-\mu)(1-p_H)]}.\end{aligned}$$

Thus

$$\begin{aligned}&\Pr[x=1|r=0,\sigma=1] - \Pr[x=1|r=0,\sigma=0] \\ &\qquad = \frac{(p_H + 3p_L - 2)(1-\mu)\mu}{(p_H - (2p_H-1)\mu - 2)(p_H + p_L - 2 - 2(p_H+p_L-1)\mu)} > 0.\end{aligned}$$

Part iv. The probability of the review decision being incorrect conditional on $r=1$ and σ is:

$$\begin{aligned}&\Pr[x=0|r=1,\sigma=1] = \frac{\Pr[p=p_H] \times [\mu(1-p_H)]}{\Pr[p=p_H] \times [\mu(1-p_H) + (1-\mu)p_H]} \\ &\Pr[x=0|r=1,\sigma=0] \\ &= \frac{\Pr[p=p_L] \times [\mu(1-p_L)] + \Pr[p=p_H] \times [\mu(1-p_H)]}{\Pr[p=p_L] \times [\mu(1-p_L) + (1-\mu)p_L] + \Pr[p=p_H] \times [\mu(1-p_H) + (1-\mu)p_H]}.\end{aligned}$$

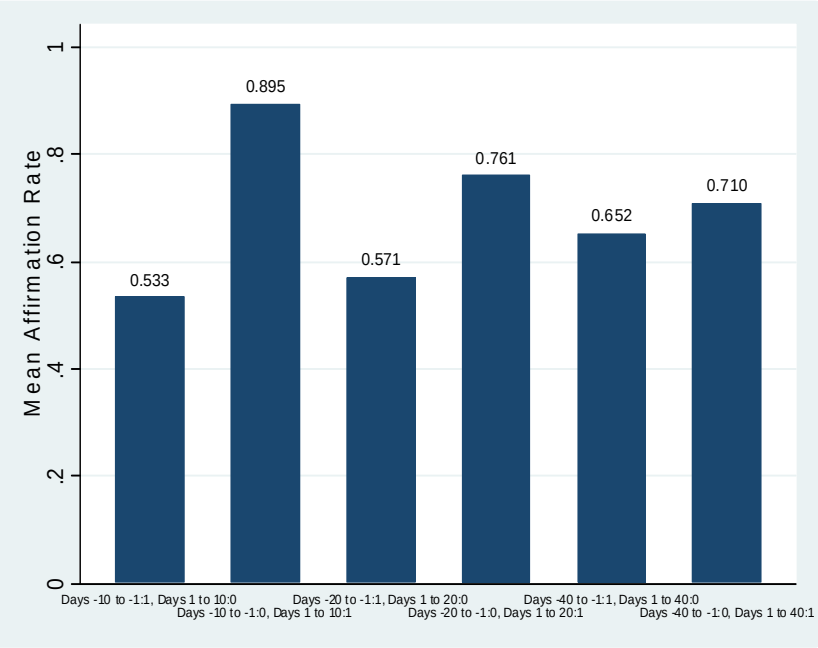
Thus

$$\begin{aligned}&\Pr\left[x=0|r=1,\sigma=1\right]-\Pr\left[x=0|r=1,\sigma=0\right]\\&\qquad=\frac{-(p_H-p_L)(1-\mu)\mu}{(p_H-(2p_H-1)\mu)(p_H+p_L-2(p_H+p_L-1)\mu)}<0.\end{aligned}$$

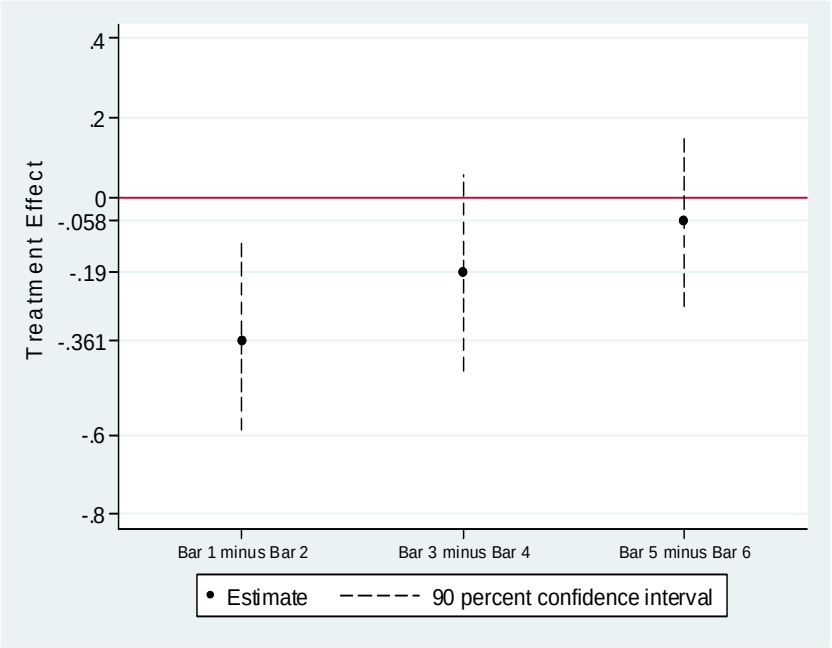
□

Figure 1— Testing for a Treatment Effect on the Review Decision

Panel A. Conditional Means

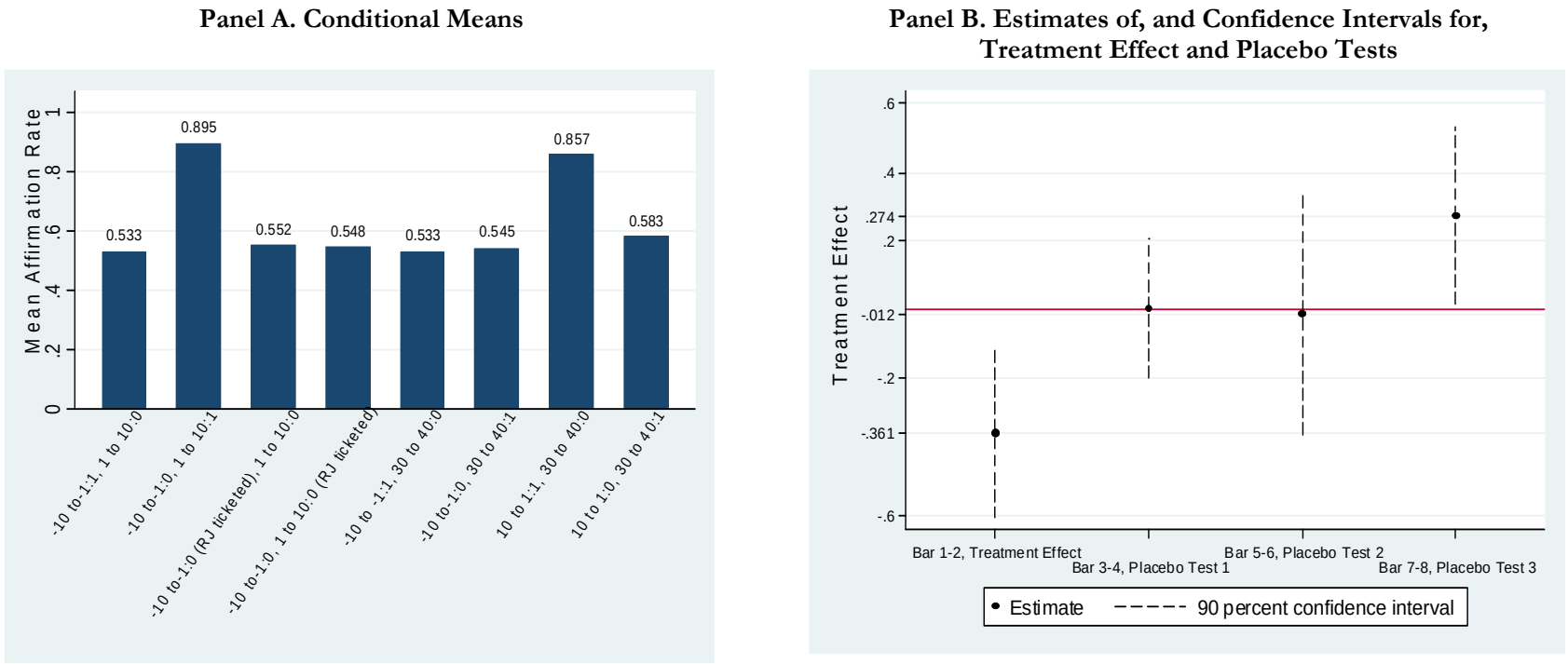


Panel B. Estimates of, and Confidence Intervals for, Treatment Effects



Notes: All bars in Figure A plot the raw data. In the first two bars in Panel A the unit of time is set at 10 days. The first bar depicts the mean affirmation rate for observations where there is one day in the 10 days before 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 immediately after the review where the panel contains the reviewed judge and one of his reviewers. The second bar depicts the mean affirmation rate for observations where there is no day in the 10 days before the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but one day in the 10 days immediately after the review where the panel contains the reviewed judge and one of his reviewers. The first plot in Panel B depicts the difference in means, -0.361, together with a 90 percent confidence interval. In the middle two bars in Panel A the unit of time is extended to 20 days. Thus, the third bar depicts the mean affirmation rate for observations where there is one day in the 20 days before the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but no day in the 20 days after the review where the panel contains the reviewed judge and one of his reviewers. The fourth bar depicts the mean affirmation rate for observations where there is no day in the 20 days before the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but one day in the 20 days after the review where the panel contains the reviewed judge and one of his reviewers. The middle plot in Panel B depicts the difference in means, -0.190, again with a 90 percent confidence interval. The final two bars in Panel A, and final plot in Panel B, repeat this exercise when the unit of time is set at 40 days.

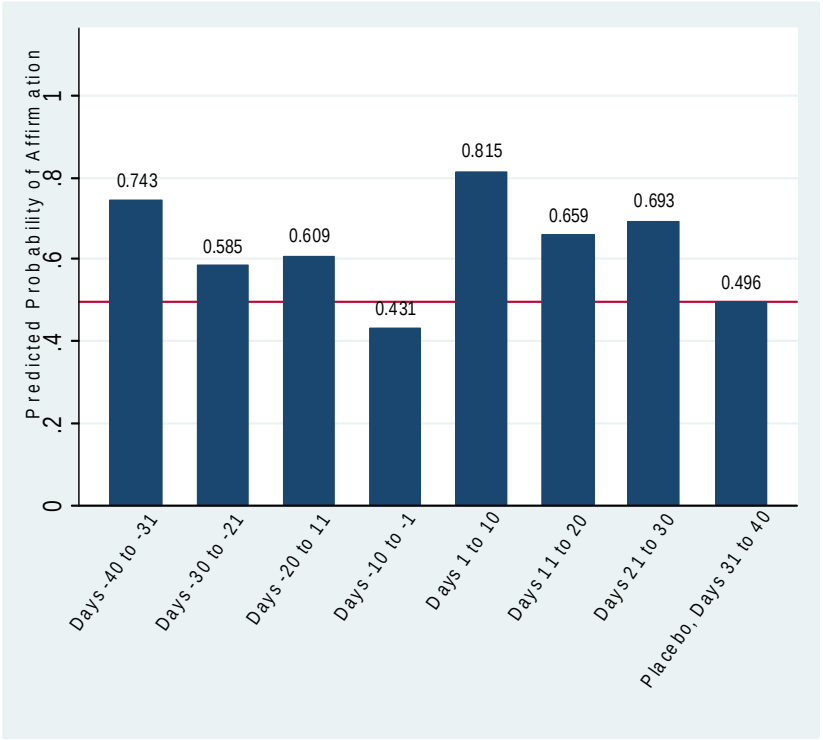
Figure 2— Placebo Tests



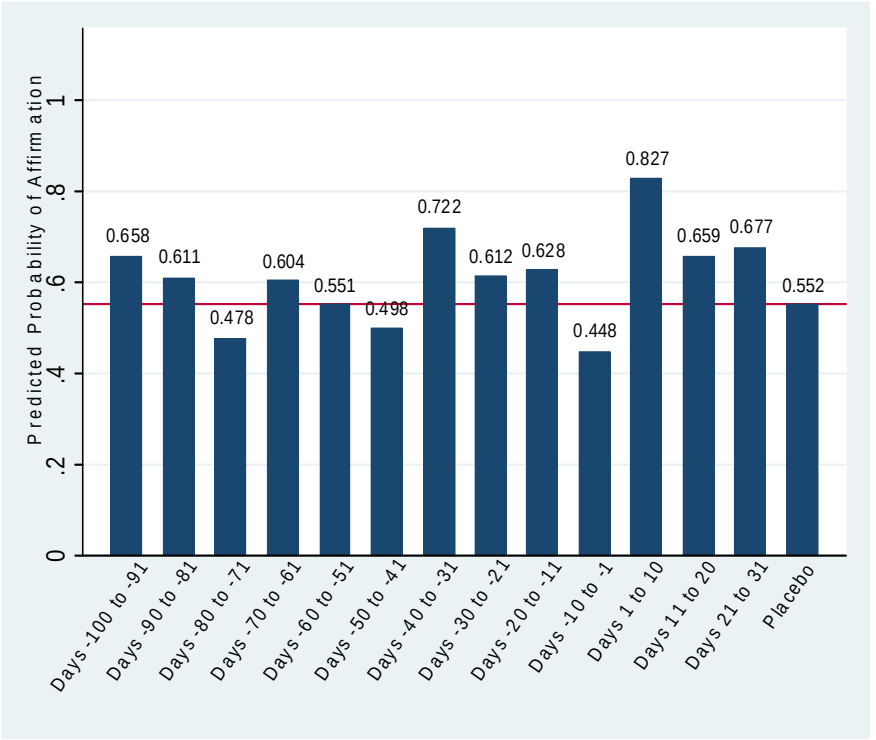
Notes: The first two bars in Panel A and the first plot in Panel B are reproduced from Figure 1. The third (fourth) bar in Panel A depicts the mean affirmation rate for observations where there is no day in the 10 days before or after the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but there is one day in the 10 days before (after), but not in the 10 days after (before), the review with a CA Civ judgement where the panel contains the reviewed judge. The second plot in Panel B depicts the difference in means, 0.004, with a 90 percent confidence interval. The fifth bar in Panel A depicts the mean affirmation rate for observations where there is one day in the 10 days before 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 sixth bar depicts the mean affirmation rate for observations where there is no treatment in the 10 days before the review but a single treatment in the 10 days *starting 30 days after* the review. The third plot in Panel B depicts the difference in means, -0.012, together with a 90 percent confidence interval. The seventh bar in Panel A depicts the mean affirmation rate for observations where there is one day in the 10 days 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 final bar depicts the mean affirmation rate for observations where there is no treatment in the 10 days immediately after the review but a single treatment in the 10 days *starting 30 days after* the review. The final plot in Panel B depicts the difference in means, 0.274, with a 90 percent confidence interval.

Figure 3— Predicted Probability of Affirmation by Time of Single Treatment

Panel A.

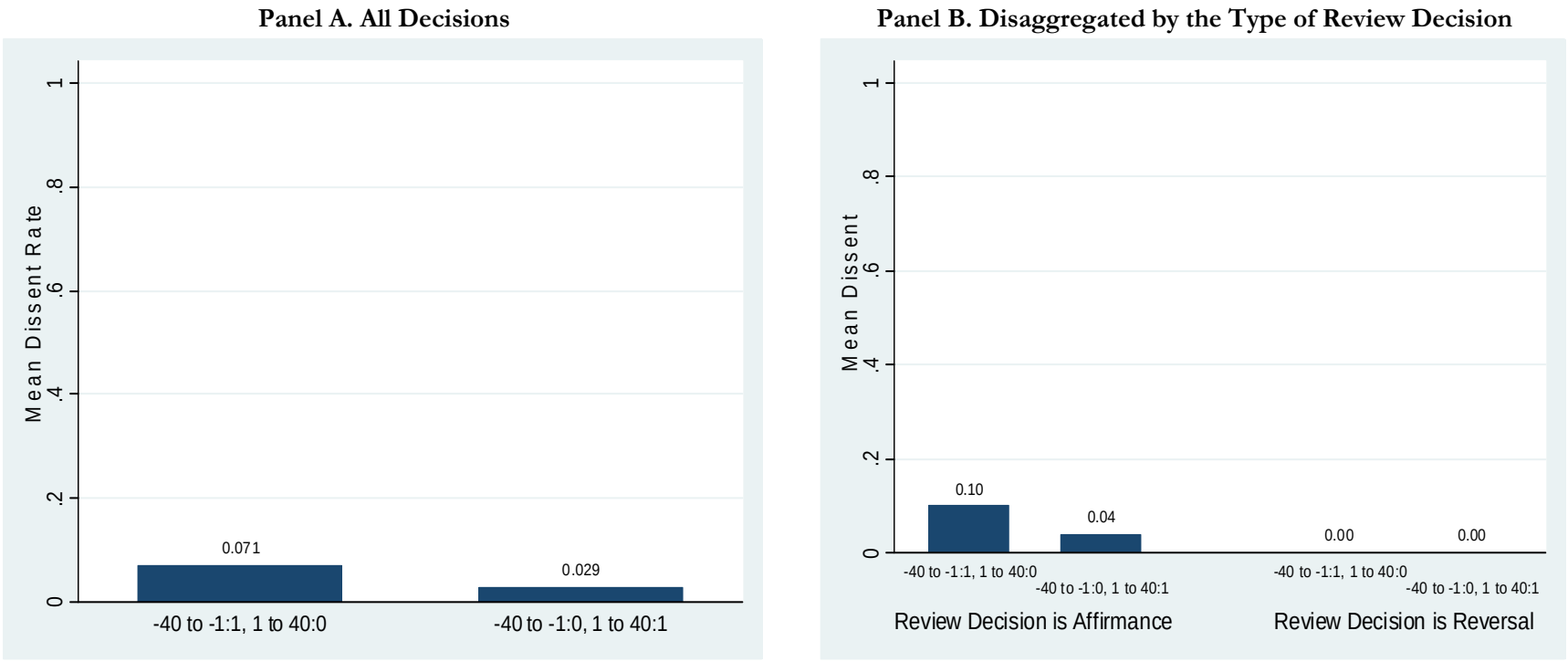


Panel B. Robustness Check Using More Periods



Notes: In Panel A, the estimates are based on the Logistic coefficients reported in column 2 of Table 5. The first bar in Panel A depicts the predicted probability of affirmation for an observation receiving a *single* treatment in the 10 day period starting 40 days before the review and *no* treatment in any other 10 day period starting within 30 days of the review (and with all other variables held at the mean for a panel with a single treatment). The next three bars represent a single treatment in the 10 day period starting, respectively, 30, 20 and 10 days before the review. The fifth bar represents a single treatment in the 10 days immediately after the review, while the final three bars represent a single treatment in the 10 day period starting respectively, 11, 21 and 31 days after the review. Panel B repeats the exercise using 100 days before and after the review. The estimates are based on the Logistic coefficients reported in column 3 of Table 5. Each bar has the same interpretation as in Panel A, except for the final bar. This bar, labelled, placebo depicts the predicted probability of affirmation for an observation receiving a single treatment in the period starting 31 days after the review and finishing 100 days after the review.

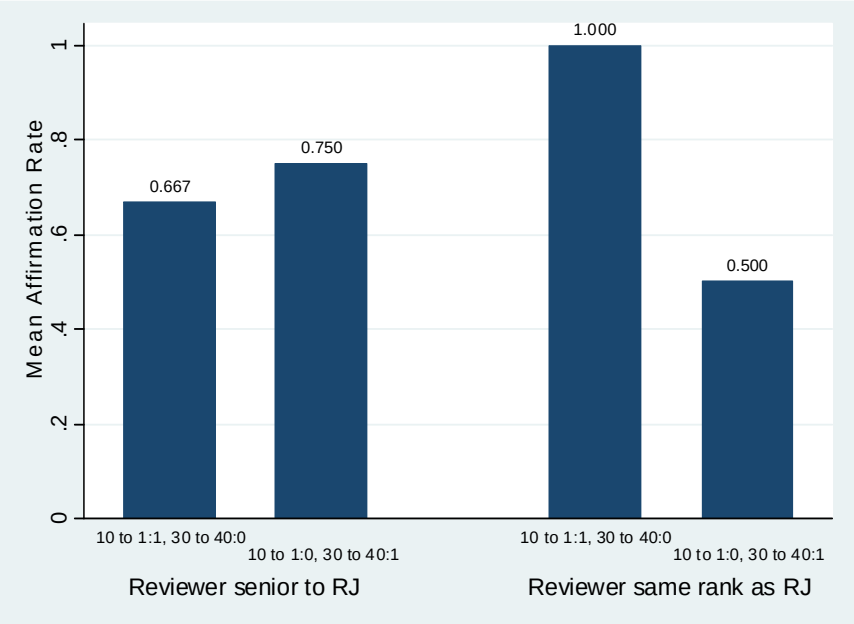
Figure 4— Dissenting Opinions During On-the-Job Interactions



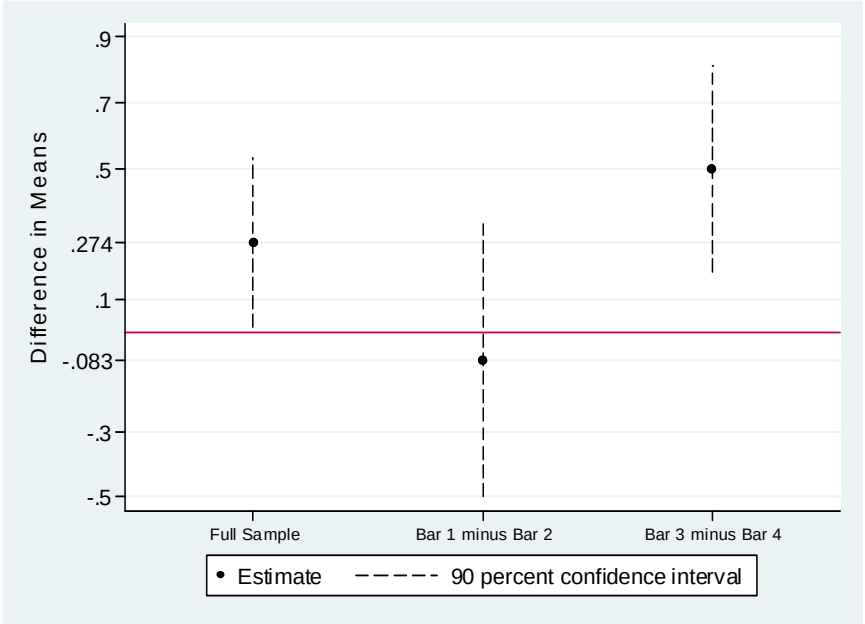
Notes: The unit of time is set at 40 days (to expand the size of the single treatment sample and hence increase the chance of observing a dissenting opinion. The first bar in Panel A depicts the mean dissent rate for observations where there is one day in the 40 days before the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but no day in the 40 days after the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers. The bar indicates that 7 percent of these pre-review interactions featured a dissenting opinion. The second bar in Panel A depicts the mean dissent rate for observations where there is one day in the 40 days 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 40 days before the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers. The bar indicates that 3 percent of these post-review interactions featured a dissenting opinion. Panel B repeats this analysis, disaggregating by the type of review decision (i.e. an affirmance or reversal).

Figure 5—Leniency by the Seniority of the Reviewer with the On-the-job Interaction

Panel A. Conditional Means



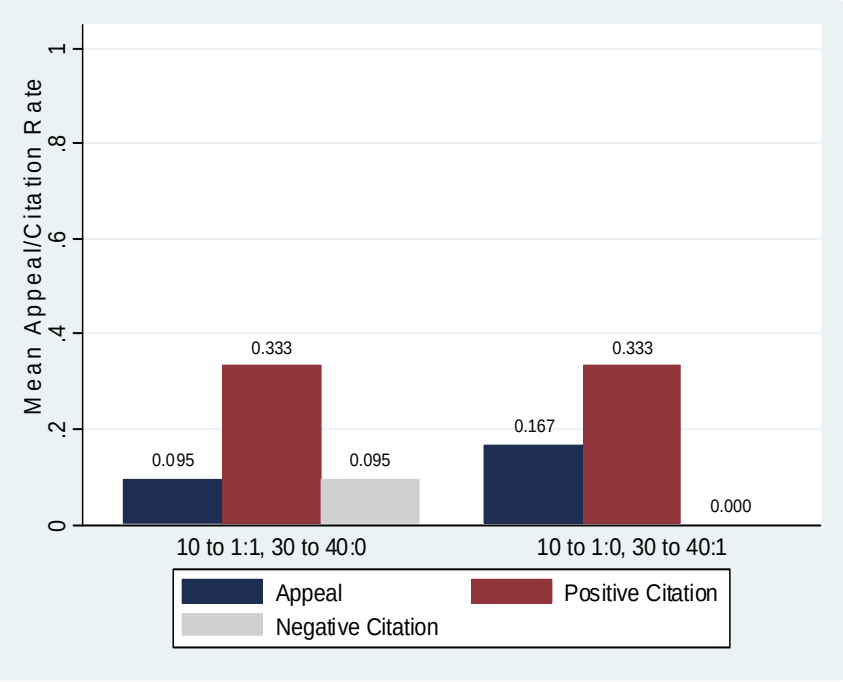
Panel B. Estimates of, and Confidence Intervals for, Treatment Effects



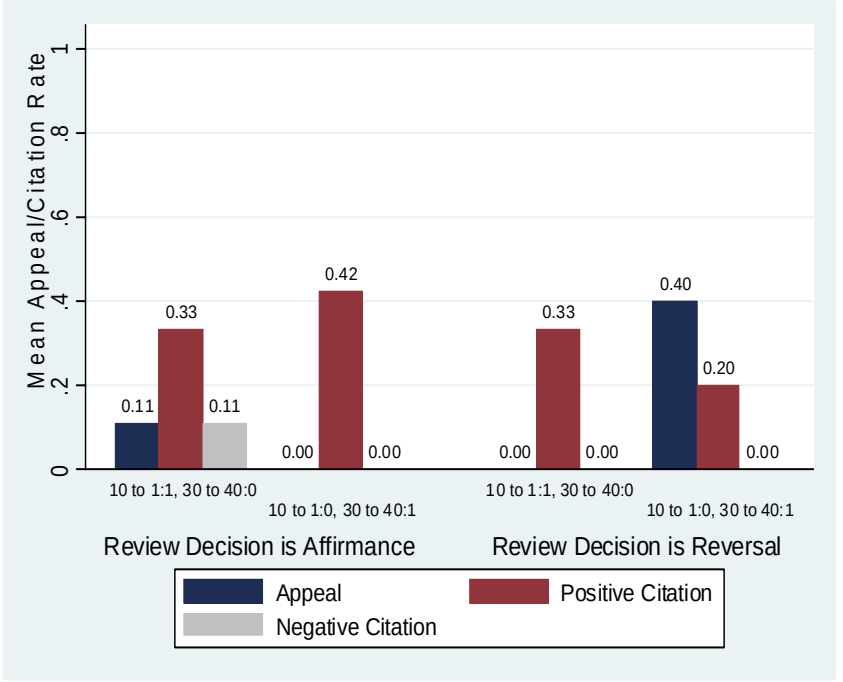
Notes: The unit of time is set at 10 days. Panel A splits the sample (from the final two bars in Figure 2 Panel A) by the relative seniority of the reviewer and reviewed judge that experience the on-the-job interaction. The first bar depicts the mean affirmation rate for observations where: (i) there is one day in the 10 days 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 and (ii) the reviewer experiencing this on-the-job interaction hold a more senior rank than the reviewed judge at the time of the review. The second bar depicts the mean affirmation rate for observations where: (i) there is no day in the 10 days immediately after the review with a CA Civ judgment where the panel contains the reviewed judge and one of his reviewers but one day in the 10 days *starting 30 days after* the review where the panel contains the reviewed judge and one of his reviewers and (ii) the reviewer with the on-the-job interaction is more senior than the reviewed judge. The first plot in Panel B depicts the difference in means for the full sample, 0.274, together with a 90 percent confidence interval, while the second plot depicts the difference in means for the 'senior' sub-sample, -0.083, with a 90 percent confidence interval. The third and fourth bars repeat this exercise for the remainder of the sample where the reviewer with the on-the-job interaction holds the same rank as the reviewed judge at the time of the review. (Note that there are no observations where the reviewer with the on-the-job interaction is *less* senior than the reviewed judge.) The final plot in Panel B depicts the difference in means for the 'same rank' sub-sample, 0.500, with a 90 percent confidence interval. The difference-in-difference estimate (second plot in Panel B minus the third plot) of -0.583 is significant at 10 percent (p=0.07).

Figure 6— The Quality of Review Decisions

Panel A. All Decisions



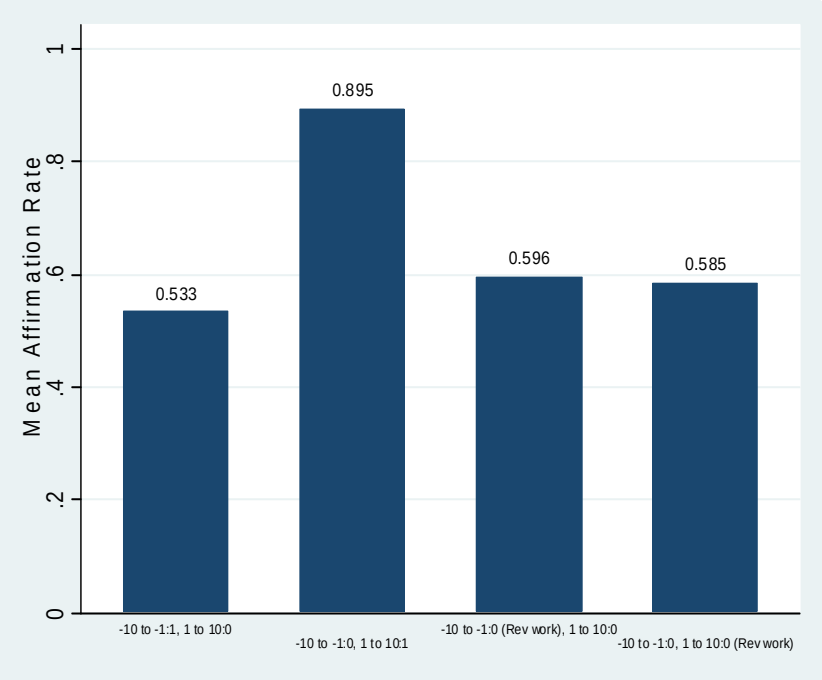
Panel B. Disaggregated by the Type of Review Decision



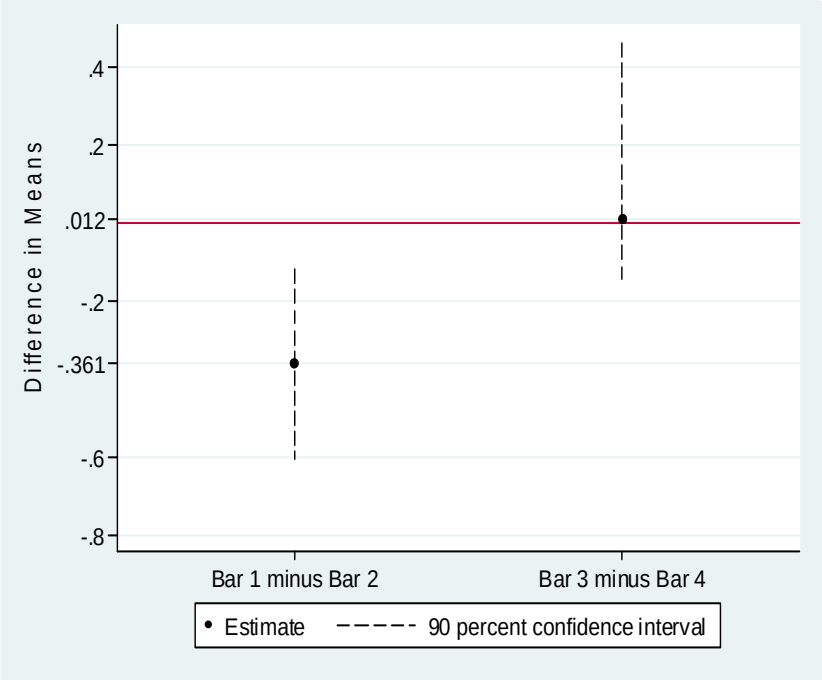
Notes: The unit of time is set at 10 days. In Panel A, the bars labelled '10 to 1:1, 30 to 40:0' depict the mean appeal (or citation) rate for observations where there is one day in the 10 days starting immediately after the review with a CA Civ judgment where the panel contains the reviewed judge and 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 '10 to 1:1, 30 to 40:0' 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).

Figure A1— Is the Timing of Treatment Effect just a Timing of Workload Effect?

Panel A. Conditional Means



Panel B. Estimates of, and Confidence Intervals for, Effects



Notes: The unit of time is set at 10 days. The first two bars in Panel A, and the first plot in Panel B, are reproduced from Figure 1. The third bar in Panel A depicts the mean affirmation rate for observations where there is one day in the 10 days before the review with a CA Civ judgment where the panel contains at least one of the reviewers *but not the reviewed judge* and no day in the 10 days immediately after the review with a CA Civ judgment where the panel contains one of the reviewers or the reviewed judge. The fourth bar in Panel A depicts the mean affirmation rate for observations where there is no day in the 10 days before the review with a CA Civ judgment where the panel contains one of the reviewers or the reviewed judge and one day in the 10 days immediately after the review with a CA Civ judgment where the panel contains one of the reviewers *but not the reviewed judge*. The second plot in Panel B depicts the difference in means, 0.012, together with a 90 percent confidence interval.

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 Unit

Panel A.

	Total treatment level $t = -1,1$	Length of Time Unit t: 5 days (1) Obs.	5 days (1) % full	10 days (2) Obs.	10 days (2) % full	20 days (3) Obs.	20 days (3) % full	40 days (4) Obs.	40 days (4) % full
Full sample	≥ 0	2262	100.0	2262	100.0	2262	100.0	2262	100.0
No treatment sample	$= 0$	2247	99.3	2221	98.2	2194	96.9	2159	95.4
Single treatment sample	$= 1$	13	0.6	34	1.5	42	1.9	54	2.4
Any treatment sample	≥ 1	15	0.7	41	1.8	68	3.01	103	4.6

Panel B.

	Length of Time Unit t: 5 days (1)	10 days (2)	20 days (3)	40 days (4)
Single treatment sample				
Proportion treated at $t = -1$	0.538	0.441	0.500	0.426
Proportion treated at $t = 1$	0.462	0.559	0.500	0.574
Difference in proportions	0.076	-0.118	0.000	-0.148
t-test for difference (*p*-value)	0.709	0.339		0.126
Any treatment sample				
Mean treatment level at $t = -1$	0.667	0.634	0.765	0.913
Mean treatment level at $t = 1$	0.467	0.561	0.838	1.078
Difference in means	0.200	0.073	-0.074	-0.165
t-test for difference (*p*-value)	0.344	0.611	0.630	0.299

Notes: Total treatment level counts the number of days in the (5, 10, 20 or 40-day) period immediately before, and the number of days in the (5, 10, 20 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 (5, 10, 20 or 40-day) period immediately before or the (5, 10, 20 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 (5, 10, 20 or 40-day) period immediately before or the (5, 10, 20 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

	Level Approach						Order Approach					
	Treated Before $S_{i,-1} > 0$ (1)		Untreated Before $S_{i,-1} = 0$ (2)		*t*-test *p*-value (3)	Norm Diff (4)	Treated Before, Untreated After $S_{i,-1} = 1, S_{i,1} = 0$ (5)		Treated After, Untreated Before $S_{i,-1} = 0, S_{i,1} = 1$ (6)		*t*-test *p*-value (7)	Norm Diff (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)	0.207	(0.405)	0.000	-0.511	0	(0)	0	(0)		0
Justice/retired	0.263	(0.444)	0.714	(0.452)	0.000	-0.712	0.348	(0.487)	0.419	(0.502)	0.602	-0.109
Above Justice	0.737	(0.444)	0.079	(0.270)	0.000	1.266	0.652	(0.487)	0.581	(0.502)	0.602	0.109
Rank reviewed judge 40 days after decision:												
Below Justice	0	(0)	0.207	(0.405)	0.000	-0.511	0	(0)	0	(0)		0
Justice/retired	0.281	(0.453)	0.710	(0.454)	0.000	-0.511	0.348	(0.487)	0.419	(0.501)	0.602	-0.109
Above Justice	0.719	(0.453)	0.083	(0.275)	0.000	0.669	0.652	(0.487)	0.581	(0.502)	0.602	-0.109
Number of reviewers	2.894	(0.451)	2.832	(0.512)	0.361	0.091	2.913	(0.417)	3	(0)	0.249	0.209
Other Observables												
Time prior ruling to review (years)	0.978	(0.615)	0.937	(0.535)	0.573	0.050	0.998	(0.840)	1.028	(0.499)	0.872	-0.030
Reviewer workload:[1]												
Total workload, *t=-1 to t=1*	25.51	(10.36)	24.16	(9.65)	0.330	0.095	26.57	(6.382)	25.07	(10.51)	0.518	0.122
Workload before the review, *t=-1*	12.98	(5.347)	11.15	(6.061)	0.024	0.227	14.09	(5.877)	11.29	(5.228)	0.071	0.356
Workload after the review, *t=1*	11.07	(6.287)	10.44	(7.116)	0.457	0.066	10.30	(7.540)	11.97	(4.491)	0.316	-0.189
Coverage of prior ruling:												
1[The Times Law Report]	0.684	(0.469)	0.331	(0.471)	0.000	0.531	0.739	(0.449)	0.387	(0.495)	0.010	0.527
1[The Independent Law Report]	0.228	(0.423)	0.117	(0.322)	0.001	0.209	0.261	(0.449)	0.065	(0.250)	0.046	0.381
No. of journal articles	3.877	(3.859)	3.022	(4.450)	0.151	0.145	3.174	(3.701)	2.806	(2.548)	0.667	0.082
Social ties with reviewed judge:												
1[At school together]	0	(0)	0.019	(0.135)	0.299	-0.116	0	(0)	0	(0)		0
1[At university together]	0.053	(0.225)	0.024	(0.152)	0.160	0.106	0.043	(0.209)	0.032	(0.180)	0.833	0.040
1[Same legal chambers]	0.175	(0.383)	0.056	(0.230)	0.000	0.266	0.217	(0.422)	0.129	(0.340)	0.399	0.162
1[Same social club]	0.140	(0.350)	0.033	(0.177)	0.000	0.273	0.130	(0.344)	0.065	(0.250)	0.418	0.153
No. of observations	57		2205		2262	2262	23		31		54	54

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 unit *t* is 40 days. Hence, *t=-1* corresponds to the 40-day period before the review (days -40 to -1).

1. 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— Robustness: Using Full Sample and Controlling for Observables

Y_i =1[Review panel *i* affirms prior ruling]	Linear Regression Models											
	Length of Time Unit *t* is 10 Days				Length of Time Unit *t* is 20 Days				Length of Time Unit *t* is 40 Days			
	(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:												
Time period -1, $S_{i,-1}$	-0.334	(0.100) ***	-0.320	(0.111) ***	-0.151	(0.070) **	-0.144	(0.073) **	-0.061	(0.043)	-0.047	(0.044)
Time period -1 and 1, $S_{i,-1} + S_{i,1}$	0.246	(0.064) ***	0.245	(0.069) ***	0.120	(0.037) ***	0.127	(0.038) ***	0.060	(0.024) **	0.066	(0.026) ***
Controls												
1[Reviewed judge below Justice at decision]			-0.078	(0.028) ***			-0.079	(0.028) ***			-0.079	(0.028) ***
1[Reviewed judge Justice/retired at decision]			Omitted				Omitted				Omitted	
1[Reviewed judge above Justice at decision]			-0.033	(0.038)			-0.040	(0.038)			-0.050	(0.039)
Number of reviewers			-0.051	(0.021) **			-0.054	(0.021) **			-0.055	(0.022) **
Time from prior ruling to review (years)			-0.010	(0.020)			-0.009	(0.020)			-0.010	(0.020)
Total workload, *t=-1* to *t=1*			0.002	(0.002)			0.001	(0.002)			0.001	(0.001)
1[The Times Law Report]			-0.016	(0.024)			-0.018	(0.024)			-0.020	(0.024)
1[The Independent Law Report]			-0.051	(0.033)			-0.049	(0.034)			-0.047	(0.034)
No. of journal articles			-0.002	(0.003)			-0.002	(0.003)			-0.002	(0.003)
1[Chancery]			0.105	(0.045) **			0.104	(0.045) **			0.101	(0.045) **
1[Civil]			0.069	(0.046) *			0.067	(0.046)			0.068	(0.046)
1[Crime]			0.117	(0.072) *			0.114	(0.072)			0.107	(0.073)
1[Employment]			Omitted				Omitted				Omitted	
1[Family]			0.132	(0.061) ***			0.136	(0.061) **			0.138	(0.061) **
1[Public]			0.147	(0.048) ***			0.146	(0.049) ***			0.147	(0.049) ***
1[At school together]			0.017	(0.077)			0.016	(0.077)			0.014	(0.076)
1[At university together]			0.039	(0.066)			0.043	(0.066)			0.039	(0.067)
1[Same legal chambers]			-0.023	(0.043)			-0.027	(0.044)			-0.026	(0.043)
1[Same social club]			-0.134	(0.057) *			-0.143	(0.056) **			-0.145	(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. In columns 1 and 2, the length of time unit is 10 days. Hence, time period -1 corresponds to the 10-day period before the review (days -10 to -1), while time period 1 corresponds to the 10-day period after the review (days 1 to 10). In columns 3 and 4, the length of time unit is 20 days. Hence, time period -1 corresponds to the 20-day period before the review (days -20 to -1), while time period 1 corresponds to the 20-days period after the review (days 1 to 20). In columns 5 and 6, the length of time unit is 40 days. Hence, time period -1 corresponds to the 40-day period before the review (days -40 to -1), while time period 1 corresponds to the 40-day period after the review (days 1 to 40).

Table 5— Robustness: Controlling for Treatment in Other Periods

Y_i =1[Review panel *i* affirms prior ruling]	Linear		Logistic			
	(1)		(2)		(3)	
Treatment level in:	Coeff	(s.e.)	Coeff	(s.e.)	Coeff	(s.e.)
Time period -10, $S_{i,-10}$					1.558	(0.692)
Time period -9, $S_{i,-9}$					1.275	(0.403)
Time period -8, $S_{i,-8}$					0.743	(0.280)
Time period -7, $S_{i,-7}$					1.236	(0.498)
Time period -6, $S_{i,-6}$					0.995	(0.557)
Time period -5, $S_{i,-5}$					0.805	(0.389)
Time period -4, $S_{i,-4}$	0.214	(0.134)	2.944	(2.136)	2.102	(1.509)
Time period -3, $S_{i,-3}$	0.077	(0.117)	1.432	(0.757)	1.280	(0.539)
Time period -2, $S_{i,-2}$	0.102	(0.106)	1.584	(0.757)	1.369	(0.530)
Time period -1, $S_{i,-1}$	-0.057	(0.118)	0.772	(0.408)	0.658	(0.291)
Time period 1, $S_{i,1}$	0.255	(0.103) **	4.478	(3.154) **	3.885	(2.593) **
Time period 2, $S_{i,2}$	0.122	(0.107)	1.963	(1.264)	1.564	(0.784)
Time period 3, $S_{i,3}$	0.153	(0.105)	2.296	(1.337)	1.696	(0.650)
Time period 4 (placebo), $S_{i,4}$	Omitted		Omitted		Omitted	
Time period 5-10 (placebo), $S_{i,t}, t = 5, \ldots, 10$					Omitted	
Total for time period -4 to 4, $\Sigma_{t=-4}^{4} S_{i,t}$	-0.054	(0.082)	0.768	(0.278)		
Total for time period -10 to 10, $\Sigma_{t=-10}^{10} S_{i,t}$					0.947	(0.080)
Controls from Table 4 included?	Yes		Yes		Yes	
Number of Observations	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 *t* is 10 days. Hence, time period -10 corresponds to the 10-day period starting 100 days before the review (days -100 to -91), time period -9 corresponds to the 10-day period starting 90 days before the review (days -90 to -81) and so forth.

Table A1— Balancing Tests, Length of Time Unit *t* is 10 days

	Level Approach						Order Approach					
	Treated Before $S_{i,-1} > 0$ (1)		Untreated Before $S_{i,-1} = 0$ (2)		*t*-test *p*-value (3)	Norm Diff (4)	Treated Before, Untreated After $S_{i,-1} = 1, S_{i,1} = 0$ (5)		Treated After, Untreated Before $S_{i,-1} = 0, S_{i,1} = 1$ (6)		*t*-test *p*-value (7)	Norm Diff (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:[1]												
Total workload, *t=-1 to t=1*	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 the review, *t=-1*	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 the review, *t=1*	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
No. 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
No. 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 unit *t* is 10 days. Hence, *t=-1* corresponds to the 10-day period before the review (days -10 to 1).

1. 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 A2— Comparison of 'Order' and 'Level' Regressions

Y_i =1[Review panel *i* affirms prior ruling]	Linear Regression Models					
	Order Regression		Level Regression (Treated Before)		Level Regression (Treated After)	
Panel A.	(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:						
Time period -1, $S_{i,-1}$	-0.334 (0.100) ***	-0.330 (0.100) ***	-0.077 (0.078)	-0.088 (0.078)		
Time period 1, $S_{i,1}$					0.239 (0.065) ***	0.225 (0.065) ***
Time period -1 and 1, $S_{i,-1} + S_{i,1}$	0.246 (0.064) ***	0.232 (0.064) ***				
Candidate Selection Variable						
1[Reviewed judge below Justice at decision]		-0.075 (0.026) ***		-0.078 (0.026) ***		-0.074 (0.026) ***
Number of Observations	2262	2262	2262	2262	2262	2262
Panel B.						
Treatment level in:						
Time period -1, $S_{i,-1}$			-0.076 (0.083)	-0.076 (0.082)		
Time period 1, $S_{i,1}$					0.264 (0.072) ***	0.260 (0.072) ***
Proxy for Ticketing Status						
Reviewed judge workload,[1] *t=-1 to t=1*			-0.001 (0.009)	-0.004 (0.009)	-0.009 (0.009)	-0.012 (0.009)
Candidate Selection Variable						
1[Reviewed judge below Justice at decision]				-0.080 (0.026) ***		-0.079 (0.026) ***
Number of Observations					2262	2262

Notes: Robust standard errors in parentheses. ***, ** and * denote significance at 1, 5 and 10 percent levels respectively. The length of time unit *t* is 10 days. Hence, time period -1 corresponds to the 10-day period before the review (days -10 to -1), while time period 1 corresponds to the 10-day period after the review (days 1 to 10).

1. This variable includes interactions. It is a count of days in the specified period (days -10 to 10) with a CA Civ judgement where the panel contains the reviewed judge and may or may not include a reviewer.