# What do Editors Maximize? Evidence from Four Leading Economics Journals*

David Card
UC Berkeley and NBER

Stefano DellaVigna
UC Berkeley and NBER

July 2017

**Abstract**

We study editorial decision-making using anonymized submission data for four leading economics journals: the *Journal of the European Economic Association*, the *Quarterly Journal of Economics*, the *Review of Economic Studies*, and the *Review of Economics and Statistics*. We match papers to the publication records of authors at the time of submission and to subsequent Google Scholar citations. To guide our analysis, we develop a benchmark model in which editors maximize the expected quality of accepted papers, and citations are unbiased measures of quality. We then generalize the model to allow different quality thresholds for different papers, systematic gaps between citations and quality, and a direct impact of publication on citations. Empirically, we find that referee recommendations are strong predictors of citations, and that editors follow the recommendations quite closely. We also show that papers by highly-published authors get more citations, conditional on the referees' recommendations and the publication status, suggesting either that referees impose a higher bar for these authors, or that prolific authors are over-cited. In their revise-and-resubmit decisions, editors only partially offset the referees' opinions, effectively discounting the citations of more prolific authors in their revise and resubmit decisions by up to 80%. We find similar patterns for the desk-reject decision. To attempt to disentangle the two explanations for this discounting, we conduct a survey of specialists, asking them to assess the relative quality of matched pairs of papers by more and less prolific authors. The respondents' assessments closely match relative citations and rule out large biases in citations as a measure of relative quality. We conclude that referees and editors at these journals set a lower bar for less prolific authors, implicitly helping younger and less established authors.

*We thank Pierre Azoulay, Esther Duflo, Patricia Funk, Matthew Gentzkow, Campbell Harvey, David Hirshleifer, Nagore Iriberri, Lawrence Katz, Imran Rasul, Laszlo Sandor, Jesse Shapiro, Vera te Velde, Ivo Welch, and Fabrizio Zilibotti for their comments and suggestions. We are also grateful to the audiences at Columbia University, MIT, UC Berkeley, UC Santa Cruz, and the Annual Meeting of WEAI for useful comments. We thank Alden Cheng, Johannes Hermle, Kaushik Krishnan, Patricia Sun, and Brian Wheaton for oustanding research assistance. We are also very appreciative of the support provided by the editors and staff at the *Journal of the European Economic Association*, the *Quarterly Journal of Economics*, the *Review of Economics and Statistics*, and the *Review of Economic Studies*. We also are grateful to the survey respondents. The survey analyzed in this paper was approved by UC Berkeley IRB, protocol 2016-08-9029 and pre-registered as trial AEARCTR-0001669.

# 1 Introduction

Editorial choices at top academic journals help shape the careers of young researchers and set the direction of research in a field. Yet, remarkably little is known about how these decisions are made. How informative are the referee recommendations that underlie the peer review process? How do editors combine the referees' advice with their own reading of a paper and other prior information in deciding whether to accept or reject it? Do editors set the same standards for established scholars as for younger or less prolific authors?

We address these questions using anonymized data on nearly 30,000 recent submissions to the *Quarterly Journal of Economics*, the *Review of Economic Studies*, the *Journal of the European Economic Association*, and the *Review of Economics and Statistics*. Our data set includes information on the field(s) of each paper, the recent publication records of the authors, whether the paper was desk rejected or sent to referees, summary recommendations of the referees, and the editorial decision of whether to reject the paper or invite it for revision. All submissions, regardless of the editor's decision, are matched to citations from Google Scholar and the Social Science Citation Index.

This unique dataset allows us to significantly advance our understanding of the editorial decision process at scientific journals. Most previous research has focused on published papers or aggregated submission data (e.g., Laband and Piette, 1994; Ellison, 2002a and 2002b; Hofmeister and Krapf, 2011; Card and DellaVigna, 2013; Brogaard, Engelberg and Parsons, 2014). While such studies offer many insights, they cannot directly shed light on the trade-offs faced by editors since they lack comprehensive information on accepted and rejected papers, including the referees' opinions. A few studies have analyzed submissions data but have focused on other issues such as the strength of agreement between referees (Welch, 2014), the effect of referee incentives (Hamermesh, 1994; Chetty, Saez, and Sandor, 2014) or the impact of blind refereeing (Blank, 1991). Two recent studies (Cherkashin et al., 2009 and Griffith, Kocherlakota, and Nevo, 2009) present broader analyses for the *Journal of International Economics* and the *Review of Economic Studies* respectively, though neither uses information on referee recommendations.

To guide our analysis we propose a simple model of the revise and resubmit (R&R) decision in which editors combine observable paper characteristics, the referee recommendations, and their own private information to select which papers to invite for revision. As a starting point we assume that editors try to maximize the expected quality of published papers and that quality is revealed by citations – i.e., citation-maximizing behavior. While highly restrictive, this stylized model is a useful benchmark for at least three reasons. First, *impact factors,* which count citations to the articles in a journal, are widely used to rank journals and are highly salient to publishers, editors and authors.[1] Second, existing studies show that citations are important determinants of career advancement (Ellison, 2012) and salaries (Hamermesh, Johnson and Weisbrod, 1982; Hilmer, Ransom and Hilmer, 2015). Finally, Google Scholar (GS) citations are available for all papers – whether accepted or rejected – including those that remain unpublished.

Nevertheless, there are at least three limitations of this simple benchmark. First, the mere

[1]There is some concern in the scientific community – e.g., Seglen (1997), Lariviere et al. (2016) – that impact factors have become too influential in making comparisons across journals.

publication of a paper in a relatively prestigious journal may raise its citations, introducing a bias in the citations as measure of paper quality. Second, editors may favor certain fields (e.g., more theoretical versus more applied fields) or certain groups of authors, effectively imposing higher and lower quality thresholds for different papers.[2] Third, even in absence of editorial preferences, citations may be systematically biased as a measure of quality by differences in citing practices across fields or a tendency to cite well-established authors (Merton, 1968).

We explicitly incorporate all three features in our modeling framework and econometric specifications. First, we allow for a direct impact of the editor's revise and resubmit (R&R) decision on ultimate citations. Using differences in R&R rates across different editors at the same journal (analogous to the so-called *judges design* used in recent studies of administrative and legal decision-making) we develop a control function that allows us to separately identify the mechanical effect of R&R on citations from the signal contained in the editor's decision. We also develop, under weaker assumptions, bounds for the impact of this mechanical effect on the key results.

Second, we allow referees and editors to express preferences (or biases) for or against certain types of papers, leading to systematic differences between the way that referee recommendations and paper characteristics are related to the accept/reject decision versus expected citations. Finally, we allow for the possibility that papers by certain authors may receive more citations, holding quality constant. Using only information on citations and editorial decisions we cannot distinguish between editorial preferences for certain types of papers and systematic gaps between citations and quality. Thus, in the final section of the paper we present data from a survey of experts in which we aim to quantify the relative bias in citations versus quality for matched pairs of papers.

We focus our main empirical analysis on the R&R decision for the just over half of submissions that are not initially desk rejected. These papers are typically reviewed by 2 to 4 referees who provide summary recommendations ranging from "Definitely Reject" to "Accept". Our first finding is that referee recommendations are strong predictors of citations: a paper unanimously classified as "Revise and Resubmit" by the referees has on average 240 log points more citations than one they unanimously agree is "Definitely Reject". We also show that the fractions of referees who rank a paper in each recommendation category provide a good summary of the information contained in the reports, with little loss relative to more flexible alternatives.

Nevertheless, a second key finding is that the referee recommendations are *not* sufficient statistics for expected citations. In particular, submissions from authors with more publications receive substantially more citations, controlling for referee recommendations. For example, papers by authors with 6 or more recent publications in a set of 35 journals have on average 100 log points more citations than papers with similar referee rankings by authors with no recent publications. This gap is essentially unchanged when we use our control function framework to measure the effect of R&R status, and is only slightly smaller under the extreme bound that treats R&R status as exogenous. We conclude that either referees are significantly tougher on more prolific authors (i.e., a bias arising

[2]Laband and Piette (1994), Medoff (2003) and Brogaard, Engelberg and Parsons (2014) all find that submissions to economics articles by authors who are professionally connected to the editor are more likely to be accepted, though they also find that papers by connected authors receive more citations, suggesting that the higher acceptance rate may be due to information rather than favoritism. Li (2017) similarly finds that members of NIH review committees tend to favor proposals in their own field, but are better informed about these proposals. In contrast, Fisman et al. (forthcoming) find strong evidence of favoritism in elections to the Chinese Academies of Engineering and Science.

from referee preferences) or that submissions from more prolific authors receive many more citations conditional on their quality (i.e., a bias in citations as a measure of quality).

Looking at the R&R decision, our third finding is that editors are heavily influenced by the referees' recommendations: the summary referee recommendations alone explain over 40 percent of the variation in the R&R decision.[3] Moreover, the relative weights that editors place on the fractions of referees in different categories are nearly proportional to their coefficients in a regression model for citations, as would be expected if editors are trying to maximize expected citations.

While the editors largely follow the referees, our fourth finding is that they also have substantial private information about the future citations of papers they handle, over and above the information contained in the summary referee recommendations (and other observable paper characteristics). In our econometric model, the reliability of this information is revealed by the coefficient on the control function that is included in our citation models to address the endogeneity of the R&R decision. Interpreted in this light, the correlation of the editor's signal with the unobserved determinants of future citations is as high as 0.20.

The R&R decision is also affected by other *observable* paper characteristics including field and author publication record, with a preference for papers from more prolific authors, conditional on the referees' evaluations. Since the referees *under-value* papers from these authors relative to expected citations, however, editors still tend to accept fewer papers from more prolific authors than would be predicted from a citation-maximizing perspective – our fifth and perhaps most surprising finding. In fact, the editors at all four journals appear to undo only a small fraction of the bias against more prolific authors exhibited by referees. This suggests either that editors agree with referees in their preference for papers from less prolific authors, or they agree with referees that the papers by more prolific authors get too many citations, conditional on their quality.

This pattern of underweighting of paper characteristics relative to the citation-maximizing benchmark is not unique to measures of the authors' publications. The editors put essentially no weight on the number of authors of a paper, despite the positive effect of a larger author team on future citations. They also do not consistently put more weight on fields with higher citations. Moreover, these patterns are not due to one or two journals: rather, they are shared by all four journals.

Although our main focus is on the R&R decision, we also analyze the desk rejection (DR) decision. Desk rejections are increasingly common in economics – accounting for about 50% of submissions in our sample – yet there is little evidence on how DR decisions are made. Our sixth finding is that editors also have substantial private information about paper quality at the DR stage. Conditional on observable characteristics including field and author publication record, papers sent for refereeing accumulate many more citations than the papers that are desk rejected. Even papers that end up rejected after refereeing have 60 log points more citations on average than papers that are desk rejected. Since both groups of papers are ultimately rejected, this comparison sidesteps any concern about endogenous publication effects. As at the R&R stage, we find that editors appear to discount the expected citations that will accrue to papers by more prolific authors at the DR stage. Indeed, desk-rejected papers by prolific authors have higher average citations than non-desk-rejected papers by authors with no previous publications.

[3] Blank (1991) and Welch (2014) similarly show that editorial decisions are highly related to the referees' opinions.

Our finding that referees *and* editors act as if they under-value citations to papers by more prolific authors runs counter to a long strand of research suggesting that the work of more prominent scholars is actually over-valued – the so-called "Matthew Effect" postulated by Merton (1968). Nevertheless, a recent review concludes that actual evidence of bias in favor of more prominent or successful authors is quite limited (Lee et al., 2013). Moreover, Blank's (1991) analysis of a randomized comparison of blind versus non-blind refereeing at the *American Economic Review* showed that blind refereeing led to *higher* relative acceptance rates for submissions from authors at top-5 schools – consistent with a bias against more prolific authors.[4] In addition, Smart and Waldfogel (1996) find *higher* citations to published articles by authors from top departments, controlling for the order of publication in the journal and page length, which they interpreted as measures of editorial treatment.[5]

To disentangle whether the discounting of citations for papers by more prolific authors arises because reviewers and editors think these papers get too many citations or because they are explicitly imposing a higher bar for these scholars, we conducted a survey of faculty and PhD students in economics, asking them to compare matched pairs of papers (published in the same year in one of the top five journals) in their field of expertise. One paper in each pair was written by author(s) with relatively many publications in the years prior to an approximate submission date, while the other was written by author(s) with few recent publications. We provided respondents with the actual Google Scholar citations for each paper and asked them to assess the appropriate relative number of citations based on their judgment of the quality of the papers. We then use the responses to infer the relative ratio of citations to quality for more versus less prolific authors, using a pre-registered specification. We emphasize that survey respondents are asked to evaluate the relative quality of the two papers, ***not*** to make R&R recommendations. Thus, we hope to abstract from any tendency to raise (or lower) the bar for more prolific authors at the refereeing stage.

Our survey respondents do not think that papers by more prolific authors receive too many citations. Indeed, their preferred relative citations for more prolific authors are only 2% below their actual relative citations (standard error = 5%), suggesting that relative citations are proportional to relative quality. In light of this finding, we argue that referees' and editors' systematic discounting of expected citations for papers written by more prolific authors arise because they impose a higher bar for these authors, leaving room in the journals for younger and less established authors.

What implications do our findings have for the editorial process at top economics journals? Overall, publication decisions are largely driven by the referees, whose summary recommendations explain 40% or more of R&R outcomes. Editors also act on some additional private information that is highly correlated with ultimate citations, though we cannot tell whether this information arises from their own reading of the paper or from the detailed referee reports (which we do not see). In either case, however, editors appear to "add value" to the decision process. On average

[4]Blank (1991, Table 10) uses information from referees on whether they knew the names of authors even when reviewing the paper under blind conditions, and constructs IV estimates of the effect of truly blind evaluation on the probability of acceptance for different groups of authors. Her results show that the acceptance rate of papers from authors at top 5 schools rises by 20 percentage points when the reviewers do not know the author's name, though the effect is imprecisely estimated.

[5]Medoff (2006) finds that papers by authors from Harvard and University of Chicago tend to receive additional citations conditional on page length and lead article status, but that authors in other top departments do not. Hofmeister and Krapf (2011) find higher citations to articles from authors at top-10 institutions, conditional on the editor's decision on which B.E. journal the paper is published in.

they also partly offset the higher bar imposed by the referees on more prolific authors, lowering the discount on their expected citations from about 100 log points to about 80 points. Editors play a more decisive role at the desk reject stage. Here, we find that their private information is also highly predictive of citations. Again, there is strong evidence of a higher bar for more prolific authors.

Given the strong competition between journals to raise impact factors, and the suspicion of many observers that the publication process is biased in favor of more accomplished authors, our conclusion that top economics journals impose a higher quality threshold for more prolific scholars is potentially surprising. Nevertheless, the similarity of behavior across the editors and reviewers at the four journals—both at the desk reject and at the R&R stage—, as well as the consistency of our results with earlier findings, including Blank's (1991) analysis of blind versus non-blind refereeing and Smart and Waldfogel's (1996) comparisons of citations for published articles, suggest that the norm of a higher bar for more prolific authors is deeply ingrained among economists. This norm likely plays a positive role in easing entry into the discipline of younger and less established authors.

# 2 Model

To help organize our empirical analysis we develop a stylized model of the editorial decision process. We first consider the R&R stage. Then we move to the earlier desk rejection stage, which shares many of the same features but with no input from the referees. For simplicity we ignore any stages after a positive R&R verdict.

## 2.1 The revise and resubmit decision

The key attribute of a paper is its quality $q$, which is only partially observed by editors and referees. At the R&R stage the editor observes a set of characteristics of the paper and the author(s), $x_1$, as well as a set of referee recommendations $x_R$.[6] Quality is determined by a simple additive model:

$$\log q = \beta_0 + \beta_1 x_1 + \beta_R x_R + \phi_q \tag{1}$$

where for simplicity we treat the unobserved component of quality, $\phi_q$, as normally distributed with mean 0 and standard deviation $\sigma_q$. Notice that we allow observable paper characteristics to help forecast paper quality conditional on the referee assessments. That is, we do not assume that the referee recommendations efficiently incorporate both the private information extracted by the referees from reading the paper and the publicly available information contained in $x_1$.[7]

[6] For simplicity in this paper we do not model the editor's decision over how many referees to assign to a paper, or the slippage between the number of referees assigned and the number who return reports. Bayar and Chemmaur (2013) discuss the optimal composition of the referee pool assigned to a given paper focusing on the role of specialist and generalist reviewers. We present some analysis below of the differences in the opinions of more and less prolific referees on the work of more or less prolific authors, which was investigated in the seminal study by Zuckerman and Merton (1971) and is related to referee "matching" (Hamermesh, 1994).

[7] It is plausible that the information contained in the recommendations varies across referees, or with the characteristics of the paper, in which case the coefficients $\beta_R$ could vary with referee characteristics or with $x_1$. We have investigated the variation in the reliability of different referees and found that this is relatively small, so for simplicity we ignore it.

The editor observes a signal $s$ which is the sum of $\phi_q$ and a normally distributed noise term $\zeta$ with standard deviation $\sigma_\zeta$:

$$s = \phi_q + \zeta.$$

Conditional on $s$ and $x \equiv (x_1, x_R)$ the editor's forecast of $\phi_q$ is:

$$E[\phi_q|s, x] = As \equiv v$$

where $A = \sigma_q^2/(\sigma_q^2 + \sigma_\zeta^2)$. This is an optimally shrunk version of the editor's private signal, and is normally distributed with standard deviation $\sigma_v = A^{1/2}\sigma_q$ and correlation $\rho_{vq} = A^{1/2}$ with $\phi_q$. The editor's expectation of the paper's quality is therefore:

$$E[\log q|s, x] = \beta_0 + \beta_1 x_1 + \beta_R x_R + v. \tag{2}$$

With this forecast in hand, the editor then decides whether to give an R&R verdict or not. Here, a natural benchmark is that the editor selects papers for which expected quality is above a threshold. Assuming $v$ has a constant variance, he or she should therefore give a positive decision ($RR = 1$) for papers with $E[\log q|s, x] \geq \tau_0$, where $\tau_0$ is a fixed threshold that depends on the target acceptance rate.[8] More generally, however, the editor may impose a threshold that varies with the characteristics of the paper or the authors. To allow this possibility we assume:

$$RR = 1 \Longleftrightarrow \beta_0 + \beta_1 x_1 + \beta_R x_R + v \geq \tau_0 + \tau_1 x_1 \tag{3}$$

where $\tau_1 = 0$ corresponds to the situation where the editor cares only about expected quality. As in a standard random preference model (McFadden, 1973) the revise and resubmit decision is deterministic as far as the editor is concerned. From the point of view of outside observers, however, randomness arises because of the realization of $s$. Under our normality assumptions, the R&R decision conditional on $x$ is described by a probit model:

$$\begin{aligned} P\left[RR = 1|x\right] &= \Phi\left[\frac{\beta_0 - \tau_0 + (\beta_1 - \tau_1)x_1 + \beta_R x_R}{\sigma_v}\right] \\ &= \Phi\left[\pi_0 + \pi_1 x_1 + \pi_R x_R\right], \end{aligned} \tag{4}$$

where $\pi_0 = (\beta_0 - \tau_0)/\sigma_v$, $\pi_1 = (\beta_1 - \tau_1)/\sigma_v$, and $\pi_R = \beta_R/\sigma_v$.

We assume that cumulative citations ($c$) to a paper, which are observed some time after the editor's decision, reflect a combination of quality and other factors summarized in a factor $\eta$[9]:

[8] Assuming that editors receive a large number of submissions and face a constraint on the total number of papers published per year, they will maximize the average quality of accepted papers by accepting a paper if and only if its expected quality exceeds some threshold $T$. If $\log q$ is normally distributed with mean $M$ and variance $V$ conditional on $(s, x)$ then expected quality is $\exp(M+V/2)$, which will exceed a given threshold $T$ if and only if $M \geq \tau_0 \equiv \log T - V/2$. We have found little evidence of heteroskedasticity in the residual from a regression of log citations on measures of $x_1$ and $x_R$, though this does not necessarily imply that $v$ is homoskedastic.

[9] As we discuss in Section 5.2, this can be easily generalized to $\log c = \theta(\log q + \eta)$, which allows a convex or concave mapping between quality and citations. Allowing $\theta \neq 1$ has no substantive effect on the implications of the model so for simplicity we set $\theta = 1$.

$$\log c = \log q + \eta.$$

The simplest possible assumption is that $\eta$ depends only on the vintage of the paper: in this case citations form a perfect index of quality, apart from an adjustment for the lag between the time the paper was evaluated and the time citations are measured. More generally, however, citations can also depend on factors like the field of a paper and the track record of the author(s) – variables included in the vector $x_1$ – as well as on the R&R decision made by the editor and other random factors captured in an error component $\phi_\eta$ :

$$\eta = \eta_0 + \eta_1 x_1 + \eta_{RR} RR + \phi_\eta. \tag{5}$$

There are two complementary hypotheses suggesting that $\eta_{RR}$ is likely to be positive. First, papers that receive an R&R verdict are likely to be published sooner than those that are rejected. This may boost citations, at least for a few years after the initial evaluation, if published papers receive more attention than unpublished works. Second, authors tend to submit their papers to higher ranked journals first, then move on to lower ranked journals if they are rejected. To the extent that publication in a higher ranked journal leads to more citations, this will also tend to boost citations for papers receiving an R&R relative to those that are rejected.

Combining equation (5) with equation (1) leads to a model for citations:

$$\begin{aligned} \log c &= \beta_0 + \eta_0 + (\beta_1 + \eta_1) x_1 + \beta_R x_R + \eta_{RR} RR + \phi_q + \phi_\eta \\ &= \lambda_0 + \lambda_1 x_1 + \lambda_R x_R + \lambda_{RR} RR + \phi \end{aligned} \tag{6}$$

where $\lambda_0 = \beta_0 + \eta_0$, $\lambda_1 = \beta_1 + \eta_1$, $\lambda_R = \beta_R$, $\lambda_{RR} = \eta_{RR}$, and $\phi = \phi_q + \phi_\eta$.

Clearly, when $\eta$ is constant across papers (and thus $\eta_1 = \eta_{RR} = 0$) we can recover $\beta_1$ and $\beta_R$ from a regression of citations on paper characteristics and referee recommendations, and potentially compare these coefficients to those estimated from the R&R probit model. More generally, however, the coefficient $\lambda_1$ in equation (6) will reflect both quality and any excess citation effect, so we cannot necessarily interpret differences in citations for papers with different observed characteristics as measures of relative quality. Moreover, OLS estimation of equation (6) poses a potential problem because $RR$ status is endogenous, and will be positively correlated with the error component $\phi$ to the extent that editors' private signals are informative about quality.

Fortunately, the structure of the editorial process suggests a straightforward approach for recovering consistent estimate of the coefficients $\lambda$ and $\lambda_{RR}$. Specifically, assume that different editors have different quality thresholds for reaching an R&R decision (i.e., different values of the constant $\tau_0$) but that the particular editor assigned to a paper has no effect on citations.[10] In this case, we can use the particular editor assigned to a paper to form an instrumental variable for R&R. We follow the approach in many recent studies of judicial and administrative decision making (e.g., Maestas, Mullen and Strand, 2013; Dahl, Kostol and Mogstad, 2014; Aizer and Doyle, 2015) and use the R&R rate on other papers handled by the same editor as a variable that shifts the threshold

[10] We defer a discussion of the validity of this assumption to Section 4 below.

for R&R but has no independent effect on citations.

Rather than estimate equation (6) by instrumental variables, we use a control function approach (Heckman and Robb, 1985; Wooldridge, 2015; Brinch, Mogstad and Wiswall, forthcoming) which in principle identifies the average treatment effect of R&R status. We first fit a probit model for the R&R decision, including $x_1$, $x_R$ and the instrumental variable $z$ formed by the leave out mean R&R rate of the specific editor. We then form an estimate of the generalized residual $r$ from the R&R probit model:

$$\begin{aligned} r &= \frac{(RR - \Phi[\pi(x,z)])\,\phi[\pi(x,z)]}{\Phi[\pi(x,z)]\,(1-\Phi[\pi(x,z)])} \\ &= \begin{cases} \frac{\phi[\pi(x,z)]}{\Phi[\pi(x,z)]} & \text{if} \quad RR=1 \\ -\frac{\phi[\pi(x,z)]}{1-\Phi[\pi(x,z)]} & \text{if} \quad RR=0 \end{cases} \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively, and

$$\pi(x,z) = \pi_0 + \pi_1 x_1 + \pi_R x_R + \pi_z z$$

is a linear index function of $x$ and $z$. Finally, we include $\widehat{r}$ (the estimate of $r$) in the citation model:

$$\log c = \lambda x + \lambda_{RR} RR + \lambda_r \widehat{r} + \phi'. \tag{7}$$

Equation (7) is a standard two-step selection-corrected model (Heckman 1976, 1979). The inclusion of the generalized residual from the R&R probit absorbs any endogeneity bias in the R&R decision. Moreover, the estimate of $\lambda_r$ provides a measure of the correlation $\rho_{v\phi}$ between the editor's private signal ($v$) and the unobserved determinants of citations ($\phi$) since $plim\ \widehat{\lambda}_r = \rho_{v\phi}\sigma_\phi$. In the special case where $\phi_\eta = 0$ (i.e., there is no additional noise in realized citations) $\rho_{v\phi} = \rho_{vq}$, and we can use the estimate of $\lambda_r$ to estimate the informativeness of the editor's signal. Otherwise, the implied correlation will tend to under-estimate $\rho_{vq}$ because citations contain an extra component of noise.

A reader might be concerned that the identity of the editor assigned to a paper affects citations, controlling for journal, field and other characteristics, or that the functional form of the control function we use is incorrect. In this case the estimated coefficients for $x_1$ and $x_R$, which are our main focus, will be potentially biased. To address these concerns we re-estimate equation (7) under two extreme assumptions which we believe bracket the plausible range of values for the coefficient $\lambda_{RR}$. For a lower bound we set $\lambda_{RR} = 0$ and estimate (7) without any control for RR status. For an upper bound we estimate (7) without including $\widehat{r}$, thereby ignoring the endogeneity of $RR$.

**Interpreting the Effects of Referee Recommendations and Paper Characteristics** In our analysis, we estimate the R&R decision model and the citation model (equations (4) and (7)) and then compare the relative effects of paper characteristics on the probability of an R&R verdict and on citations. As a starting point, consider a model with two simplifying assumptions:

- (A1) the editor only cares about quality ($\tau_1 = 0$)
- (A2) citations are unbiased measures of quality ($\eta_1 = 0$).

Notice that we are not assuming that the referee recommendations are unbiased predictors of a paper's quality. If referees tend to give worse (better) recommendations to certain types of papers, controlling for quality, this will lead to a positive (negative) coefficient for the corresponding element of $x_1$ in equation (7). An estimate of 0.10 for the coefficient of a certain paper characteristic in equation (7), for example, means that the referees are essentially discounting expected citations to papers in this category by 10% in making their recommendations about publishing the paper.

Assuming that the editor is a citation maximizer, however, he or she will take the referees' biases into account and weight papers with this characteristic more positively. Specifically, under assumptions A1 and A2 the editor will use weights in the R&R decision rule that are strictly proportional to the weights that the referee reports and the paper characteristics receive in the citation model, leading to the prediction:

(P1) $\pi_1 = \lambda_1/\sigma_v$, $\pi_R = \lambda_R/\sigma_v$.

Figure 1 illustrates the testable implications of this prediction. If we graph the estimated coefficients $(\widehat{\pi}_1, \widehat{\pi}_R)$ from the R&R probit against the corresponding estimated coefficients $(\widehat{\lambda}_1, \widehat{\lambda}_R)$ from the model for log citations, the points should lie on a positively sloped line that passes through the origin with slope $1/\sigma_v$. As we show, these restrictions are not fully satisfied at any of the four journals in our sample, leading us to consider the sources of the violations.

Dropping either A1 or A2 allows for systematic departures between the relative effect of $x_1$ and $x_R$ on the probability of an R&R versus observed citations. In either case the referee recommendation variables will still affect citations and the R&R decision proportionally, so the coefficients $\widehat{\pi}_R$ and $\widehat{\lambda}_R$ will continue to lie on a positively sloped line with slope $1/\sigma_v$. Now, however, the coefficients of the $x_1$ variables may lie above or below this line. For a characteristic that leads the editor to impose a higher (lower) R&R threshold, the corresponding pair of coefficients $(\widehat{\pi}_{1k}, \widehat{\lambda}_{1k})$ will fall below (above) the reference line plotting the $\widehat{\pi}_R$ coefficients against the $\widehat{\lambda}_R$ coefficients. Similarly, for a paper characteristic that leads to more (less) citations conditional on the paper quality, the corresponding pair of coefficients will fall below (above) the reference line. The two alternative explanations for any non-proportional effects can only be distinguished if we measure the relationship between quality and citations, which our survey of expert readers described below attempts to uncover.

Regardless of the source of non-proportionality, we can devise a simple metric based on log citations that summarizes the degree of non-proportionality. Specifically, consider the $k^{th}$element of the vector $x_1$. Let $\pi_{1k}$ represent the coefficient of this characteristic in the R&R probit model. If $\tau_{1k} = \eta_{1k} = 0$ then the coefficient of this characteristic in the log citation model would be:

$$\lambda^*_{1k} = \pi_{1k}\sigma_v$$

We estimate $\lambda^*_{1k}$ using the estimate of $\pi_{1k}$ from the probit model and an estimate of $\sigma_v$ based on the relative effect of the referee recommendations on citations and the R&R decision. The gap $(\lambda_{1k} - \lambda^*_{1k})$ between the actual coefficient in the citation model and the predicted coefficient with $\tau_{1k} = \eta_{1k} = 0$ represents the extent to which citations to papers with the $k^{th}$ characteristic are over- or undervalued by the editor. Notice that if $\pi_{1k} = 0$ then $\lambda^*_{1k} = 0$: the editor agrees with the referees and applies the same discount factor. But for characteristics with a positive (negative) coefficient

in the R&R probit model, the editor is applying a smaller (larger) discount than the referees.

## 2.2 The Desk Reject Decision

Having analyzed the R&R decision, we now consider the earlier desk rejection decision. At this stage the only observable information is $x_1$. We assume that conditional on $x_1$ paper quality is

$$\log q = \alpha_0 + \alpha_1 x_1 + \omega_q \tag{8}$$

where $\omega_q$ is a normally distributed error component with mean 0 and standard deviation $\sigma_{\omega_q}$. Based on an initial reading of the paper the editor observes a signal

$$s_0 = \omega_q + \varepsilon$$

where $\varepsilon$ is normally distributed with mean 0 and standard deviation $\sigma_\varepsilon$. Conditional on this information, the editor's estimate of the expected quality of the paper is

$$\begin{aligned} E[\log q|x_1, s_0] &= \alpha_0 + \alpha_1 x_1 + A_0 s_0 \\ \text{where} \quad A_0 &= \frac{\sigma^2_{\omega_q}}{\sigma^2_{\omega_q} + \sigma^2_\varepsilon}. \end{aligned}$$

Define $v_0 = A_0 s_0$: this is a normally distributed random variable with mean 0 and standard deviation $\sigma_{v_0} = A_0^{1/2}\sigma_{\omega_q}$ that is observed by the editor but is unknown to outside observers. We assume the editor assigns a paper for review (i.e., does not desk reject the paper) if

$$E[\log q|x_1, s_0] = \alpha_0 + \alpha_1 x_1 + v_0 \geq \gamma_0 + \gamma_1 x_1$$

which has the same form as the decision rule at the R&R stage.[11] This rule leads to a simple probit model for the probability of non-desk-rejection ($NDR = 1$), conditional on the characteristics $x_1$:

$$P[NDR = 1|x_1] = \Phi[\frac{\alpha_0 - \gamma_0 + (\alpha_1 - \gamma_1)x_1}{\sigma_{v_0}}]. \tag{9}$$

Next we specify a model for citations, conditional on information available at the desk reject stage. We assume that the gap between citations and quality, conditional on $x_1$ and NDR status, can be written as:

$$\log c - \log q = \delta_0 + \delta_1 x_1 + \delta_{NDR} NDR + \omega_c. \tag{10}$$

[11] An optimal desk reject rule compares the option value of refereeing a paper to the cost of refereeing. Assume as in equation (3) that the R&R decision rule compares the conditional expectation of log quality, given $x_1$, $x_R$ and a later signal $s$ to some threshold $\tau(x_1)$. Then the optimal rule for not desk rejecting ($NDR$) is

$$NDR \Leftrightarrow \int\int max[0, E[q|x_1, s_0, x_R, s] - \tau(x_1)]f(x_R, s|x_1, s_0)dx_R ds - C > 0$$

where $f(x_R, s|x_1, s_0)$ is the joint density of $(x_R, s)$ conditional on the information observed at the desk reject stage, and $C$ is the cost of refereeing. We assume this can be approximated by a cutoff rule of the form $E[\log q|x_1, s_0] > \gamma_0 + \gamma_1 x_1$.

which includes a constant, a random error component, and potential controls for $x_1$ and NDR status. As we argued at the R&R stage, it is plausible that $\delta_{NDR}$ is positive. Non-desk-rejected papers have some chance of being published in the current journal, whereas those that are desk rejected have to be submitted to other outlets. Other things equal, NDR papers are likely to be published sooner and to be published in a higher quality journal – both factors that could raise citations.

Combining equations (10) and (8) leads to a model for observed citations conditional on $x_1$:

$$\begin{aligned} \log c &= \alpha_0 + \delta_0 + (\alpha_1 + \delta_1)x_1 + \delta_{NDR}NDR + \omega \\ &= \psi_0 + \psi_1 x_1 + \psi_{NDR}NDR + \omega \end{aligned} \tag{11}$$

where $\psi_0 = \alpha_0 + \delta_0$, $\psi_1 = \alpha_1 + \delta_1$, $\psi_{NDR} = \delta_{NDR}$, and $\omega = \omega_q + \omega_c$. Since desk rejection is determined in part by the editor's signal, we expect that the error term will be positively correlated with $NDR$. As at the R&R stage, we address this using a control function approach. We first fit a probit model for NDR, including the observable paper characteristics and an instrumental variable $z_0$ based on the mean NDR rate of the editor on other papers. We then form an estimate of the generalized residual from this probit model, $r_0$, and estimate a selection-corrected citation model:

$$\log c = \psi_0 + \psi_1 x_1 + \psi_{NDR}NDR + \psi_r \widehat{r}_0 + \omega'.$$

The coefficient $\psi_r$ captures the strength of the correlation between the residual in the NDR probit, which is based on the editor's signal, and the residual $\omega = \omega_q + \omega_c$. This will be larger the better able is the editor to forecast quality, and the stronger the link between citations and quality (i.e., the smaller is the variance of the noise component $\omega_c$).

**Comparisons of Papers with the Same Probability of Desk Rejection** At the desk rejection stage we do not have a set of variables, comparable to the referee recommendations, that provide a ready benchmark for gauging the effects of a given characteristic on expected citations and the probability of desk rejection. Nevertheless, it is possible to develop a simple test of citation-maximizing choice behavior based on comparisons of citations for non-desk-rejected papers with the same probability of NDR.[12] Specifically, our model implies that expected citations should be similar for any two papers with the same probability of non-desk-rejection handled by a given editor.

To see why, suppose that $\gamma_1 = \delta_1 = 0$. In this case that the probability of non-desk rejection by a given editor is:

$$p(x_1) = P[NDR = 1|x_1] = \Phi[\frac{\alpha_0 - \gamma_0}{\sigma_{v_0}} + \frac{\alpha_1}{\sigma_{v_0}}x_1]$$

where the editor's NDR threshold is captured by the constant $\alpha_0$. Expected citations for a paper that is NDR by this editor are :

$$E\left[\log c|x_1, NDR = 1\right] = \alpha_0 + \delta_0 + \alpha_1 x_1 + \psi_{NDR} + \psi_r g(\frac{\alpha_0 - \gamma_0}{\sigma_{v_0}} + \frac{\alpha_1}{\sigma_{v_0}}x_1)$$

[12]This test is similar to the tests widely used in the law and economics literature to test for discrimination by police officers in deciding to stop people of different race groups, e.g., Knowles, Persico and Todd (2001).

where $g(y) = \frac{\phi[y]}{\Phi[y]}$ is the standard "selection correction" term specified in Heckman (1979). Now consider any two papers that receive an NDR verdict from a given editor with the same value for $p(x_1)$. These papers must have the same value for the covariate index $\alpha_1 x_1$, and thus the same expected citations. The assumption of citation maximizing behavior therefore implies:

$$E\left[\log c | x_1, NDR = 1\right] = G(p(x_1)) \tag{12}$$

where $G(\cdot)$ is a strictly increasing continuous function.

Equation (12) leads to a simple, intuitive test for the citation maximizing hypothesis: we fit a model for the probability of NDR, then classify papers into cells based on their propensity to receive an NDR verdict, and compare average citations for papers handled by a given editor with different values of an individual covariate (such as the author's previous publication record). Under citation maximization, $p(x_1)$ is a sufficient statistic for expected citations among all papers with NDR=1, and there should be no difference in expected citations for papers in a given cell. If, on the other hand, editors are using a different threshold for different authors (i.e., $\gamma_1 \neq 0$) or editors care about quality but citations are a biased measure of quality (i.e., $\delta_1 \neq 0$) then we expect to see differences in expected citations for papers with the same NDR propensity.

# 3 Data

**Data Assembly.** We obtained permission from the four journals in our sample to assemble an anonymized data set of submissions that for each paper combines information on the year of submission, approximate field (based on JEL codes at submission), the number of co-authors and their recent publication records, the summary recommendations of each referee (if the paper was reviewed), an (anonymized) identifier for the editor handling the paper[13], citation information from Google Scholar (GS) and the Social Science Citation Index (SSCI), and the editor's decisions regarding desk rejection and R&R status.[14]

Our data assembly process relies on the fact that all four journals use the Editorial Express (EE) software system, which stores information about past submissions in a set of standardized files that can be accessed by a user with managing editor permissions. We wrote a program that extracted information from the EE files, queried the GS system, and merged publication histories for each author from a data base of publications in major journals (described below). The program was designed to run on a stand-alone computer under the supervision of an editorial assistant and create an anonymized output file that is stripped of all identifying information, including paper titles, author names, referee names, and exact submission dates. For additional protection the citation counts and publication records of authors are also top-coded.[15] We constructed our data sets for

[13] As per our agreement with the journal, we did not store editor identifiers for REStat. In the analysis with editor fixed effects, we treat REStat as having just one editor. In that analysis, we also pool, within a journal, editors who handled very few papers, as the mean R&R rate would be very imprecisely estimated.

[14] The data set does not include any information on demographic features of the authors or referees, such as age or gender, and does not track authors or referees across papers.

[15] The top-code limit for citations is lower for REStud than the other journals. We adjust for this using an imputation procedure based on the mean of citations at the other journals for papers that are above the REStud topcode.

the *Review of Economics and Statistics* (REStat) and the *Quarterly Journal of Economics* (QJE) in April 2015, and the data set for the *Review of Economic Studies* (REStud) in September 2015. The data set for the *Journal of the European Economic Association* (JEEA) was constructed over several months up to and including September 2015.

**Summary Statistics.** We have information on all new submissions (i.e., excluding revisions) to each of the four journals from their date of adoption of the EE system until the end of 2013, allowing at least 16 months for the accrual of citations before citations are measured. As shown in Table 1, we have data beginning in 2005 for the QJE (N=10,824) and REStud (N=8,335), beginning in 2006 for REStat (N=5,767), and beginning in 2003 for JEEA (N=4,942).

Table 1 and Figure 2a present information on the editorial decisions for the papers in our sample. Desk rejections are more common at the QJE and REStat (60% and 54% of initial submissions respectively) than at REStud or JEEA (20% and 24%, respectively). The R&R rate is lowest at the QJE (4%) and highest at REStat (12%). We do not keep track of the revision stages that occur after an initial R&R decision.[16]

Figure 2b and Columns 6-10 of Table 1 provide information on a key input to the editorial process: the referee recommendations for papers that are not desk-rejected. The EE system allows referees to enter one of 8 summary recommendations ranging from "definitely reject" to "accept".[17] The modal recommendation is "reject" at all four journals; a majority of recommendations (ranging from 54% at REStat to 73% at QJE) are "definitely reject" or "reject".[18]

We use the JEL codes provided by the author(s) to determine whether the paper belongs to one of 15 field categories listed in Table 1. To account for multiple field codes we set the indicator for a field equal to 1/J where J is the total number of fields to which the paper is assigned. The most common fields are labor, macro, and micro. The field distributions vary somewhat across journal, with a higher share of theory submissions at REStud and a higher share of labor economics at QJE.

An important variable is the publication record of the author(s) at the time of submission. To code this variable, we extracted all articles published in 35 high-quality journals between 1991 and 2014. The set of journals (shown in Appendix Table 1) includes the leading general interest journals as well as top field journals in a majority of fields. We construct the total number of papers published by a given author in these journals in a 5-year window ending in each year from 1995 to 2013[19]. We then take the highest publication record of all co-authors, setting the count to 0 if we find no previous publications. For example, a paper written by a team in which the most prolific coauthor published 4 papers in the 35 journals in the 5 years up to and including the year of submission is coded as having 4 papers. We also keep track of the number of coauthors, since this is a positive predictor of citations among published papers (Card and DellaVigna, 2013).

As shown in Table 1 and Figure 2c, 46% of submissions in our overall sample were submitted

[16]We do have information on final publication status for REStud and JEEA. Among papers submitted up to 2011 the final publication rate for papers that received a positive R&R verdict was approximately 90% at JEEA and 75% at REStud.

[17]The top two categories are "conditionally accept" and "accept". Since these recommendations are rare, we pool both under the accept category.

[18]Welch (2014, Table 3) shows the distributions of referee recommendations at 6 economics journals (including the QJE and 5 others) and 2 finance journals. These distributions are quite similar to the ones in our data.

[19]We also store the publications in these same journals in years 6-10 before submission, and the number of publications in the top 5 economics journals in the 5 years before submission.

by authors with no previous publications (or whose names could not be matched to our publication database), while 17% were submitted by authors with 4 or more publications. Submissions at the QJE tend to come from the most prolific authors, followed by REStud, then REStat and JEEA.

A final key piece of information is the number of citations received by a paper. We recorded citations as of April 2015 for QJE and REStat and as of August 2015 for REStud and JEEA. For our main measure we use GS, which provides information regardless of whether a manuscript is published or not. This is particularly important in our context because we are measuring citations for some of the papers in our sample only 2-3 years after the paper was submitted, and we want to minimize any mechanical bias arising because papers that are rejected take some time to be published in other outlets, or may never be published. As a robustness check, we also use counts of citations from the SSCI, which are reported in GS but are only available for published papers (and only count citations in other published works).

We merge citation counts to papers using the following procedure. First we extract a paper's title from EE and query GS using the *allintitle* function, which requires all words in the EE title to be contained in the GS title. We capture the top 10 entries found under the allintitle search, and verify that a given GS entry has at least one author surname in common with the names of authors in EE. Then the GS and SSCI citation counts for all entries with a matching name are summed to determine total citations. Thus, we add the citations accrued in working paper format and in the final publication, as long as the paper title is the same. Papers with no match in Google Scholar are coded as having zero citations.[20]

Working with citations raises two issues. First, citation counts are highly skewed: about 30% of submitted papers have no citations, with an even higher rate among recent submissions. Second, citations to a given paper rise with the passage of time. We use two complementary approaches to address these issues. For our main specifications we use the inverse hyperbolic sine (asinh) of the citation count and include journal-year fixed effects. The asinh function closely parallels the natural logarithm function when there are 2+ citations, but is well defined at 0.[21] Online Appendix Figure 1a shows the distribution of $asinh(citations)$ in our sample, with a spike at 0 (corresponding to 30% of papers with 0 cites) and another mode at around 3 (corresponding to around 10 cites). Under this specification, we can interpret the coefficients of our models as proportional effects relative to submissions from the same journal-year cohort (i.e., as measuring log point effects). As an alternative we assign each paper its citation percentile within the pool of papers submitted to the same journal in the same year. To eliminate heaping we randomly perturb the number of citations received by each paper, smoothing out the 30% of papers with 0 citations (see Online Appendix Figure 1b).

[20] In Online Appendix Table 8 we show that our main results are robust to an alternative choice in which papers with no match in GS are dropped from the analysis.

[21] Asinh(z)=$\ln(z + \sqrt{1 + z^2})$. For z$\geq$ 2, asinh(z)$\approx$ ln(z) + ln(2), but asinh(0)=0.

# 4 Empirical Results

## 4.1 Models for Citations and The R&R Decision

### Summarizing Referee Opinions

How informative are referee recommendations about future citations? We consider the 15,177 papers that were not desk-rejected and were assigned to at least two referees. This choice reflects the fact that in many cases assignment to a single referee is equivalent to desk rejection. In particular, papers at REStud assigned to only one referee have a 99% rejection rate. We therefore exclude the 2,271 papers assigned to one referee, though the estimated coefficients in our main models are very similar regardless of whether we include or exclude these papers at all journals or only at REStud.

Figures 3a and 3b show how citations are related to referee recommendations. To construct these figures we take each paper/referee combination as an observation and calculate mean citations by the referee's summary recommendation, weighting observations by the inverse of the number of referee recommendations for the associated paper. Figure 3a uses asinh of the number of citations, while Figure 3b uses the citation percentile within the same journal×year submission cohort.

Both figures show a clear association between referee recommendations and citations, though the effect is somewhat nonlinear, with a relatively large jump between *Definitely Reject* and *Reject,* and a negligible change between *Strong Revise and Resubmit* and *Accept*. The slope of the relationship is quite similar across journals, suggesting a similar informativeness of referees across journals. The *levels* of the citation measures differ, however, with the highest citation levels at the QJE and the lowest at JEEA. The differences in the citation percentile measures are driven by differences in the degree of selectivity of the papers that are reviewed relative to the overall submission pool at each journal. This selection process is strongest at the QJE, where only 40% of papers are reviewed and the average citation citation percentile for all reviewed papers is 65, and weakest at JEEA, where about 65% of papers are reviewed and the average citation percentile of these papers is 53.[22]

Figures 3a and 3b relate mean citations to the opinions of individual referees. How do citations vary with the collective opinions of the entire team of referees? Figure 3c presents a heat map of mean citations for papers with 2 reports, showing the data for each of the 7×7=49 possible cells for the two referee's recommendations.[23] The figure reveals that average citations depend on the average opinions of the referees. For example, papers receiving two *Reject* recommendations have a mean $asinh(citations)$ of 2.5, while papers with two *Strong R&R* recommendations have a mean of 4.1. Papers with one *Reject* and one *Strong R&R* fall in the middle with a mean of 3.2. In Online Appendix Figure 2c we present parallel evidence for papers with 3 reports, creating a heat map using all possible pairs of recommendations. These data support the same conclusions.

In light of this evidence, we summarize the referee recommendations using the fractions of recommendations for a given paper in each of the 7 categories. For example, if a paper with 3 reports

[22] With 40% of papers reviewed, the expected citation percentile if the desk rejection process perfectly eliminated the bottom tail is 80, while with a 65% review rate the expected citation percentile under perfect selection is 67.5. Using these as benchmarks, the efficiency of the desk rejection is 65/80=0.81 at QJE and 53/67.5=0.79 at JEEA.

[23] The referees' recommendations are modestly positively correlated, with rank order correlations of around 0.25 for 2-referee papers. Welch (2014) shows similar correlations for referee recommendations at a broader sample of economics and finance journals.

has two referees recommending *Reject* and one referee recommending *Weak R&R* then the fractions are 2/3 for *Reject,* 1/3 for *Weak R&R* and 0 for all other categories. This simple procedure has the benefit that it can be used irrespective of the number of reports.

Column 1 in Table 2 reports the estimates of an OLS regression model for $asinh(citations)$ that includes journal × year fixed effects and the fractions of referee reports in each category. As in the figures, the estimates suggest a strong positive effect of referee enthusiasm on mean citations. The increases in the estimated coefficients between categories are substantially larger than the slopes in Figure 3a, reflecting the fact that the coefficients in the table reflect the effect of all referees unanimously changing recommendations from one category to another, whereas the figure reflects the effect of only one referee changing his/her recommendation.

To document the validity of our averaging specification we return to the subsample of papers with two reports, and display in Figure 3d the predicted citations from the model in column 1 of Table 2 in each of the 49 cells. Comparing these predictions with the actual citations in Figure 3c shows that the model does a very good job of summarizing the recommendations. The model also does well for papers with 3 reports, as shown by comparing Online Appendix Figures 2c and 2d. Moreover, as shown in Online Appendix Table 4 when we compare the coefficients of the referee category variables for papers with 2, 3, and at least 4 referees, the coefficients are remarkably similar.

**Other Determinants of Citations**

Next we consider other possible determinants of citations, including the recent publication record of the authors, the number of authors and the field of the paper. Without controlling for referee recommendations, these variables are strong predictors of citations (column 2 of Table 2). An increase in the number of author publications from 0 to 4 or 5, for example, raises citations by about 100 log points, a large (and highly statistically significant) effect. The effect of the number of authors is not as large, though still sizable (and highly significant). Relative to a single-authored paper (the base category) a paper with 3 co-authors has 24 log points more citations (roughly 27% more). There are also systematic differences in citations for different fields (see Online Appendix Table 2): papers in theory and econometrics have the lowest citations, while papers in international and experimental economics have the highest citations. These differences are broadly consistent with patterns in the existing literature based on published papers (e.g., Card and DellaVigna, 2013).

To what extent do these effects persist after controlling for referee recommendations? As noted in Section 2, if the referee reports are a sufficient statistic for quality, and citations are unbiased measures of quality, then the other covariates should have no effect on citations after controlling for the referees' recommendations. Within the framework of our model, variables that remain significant predictors of citations indicate that the referees either believe that citations should be discounted for certain groups to properly measure quality, or that certain types of papers should be more highly rated holding constant their quality.

Column 3 in Table 2 presents a specification with both referee recommendations ($x_R$ in our notation) and the other controls ($x_1$). The referee variables remain highly significant predictors, with coefficients attenuated by about 15 percent relative to the specification with no controls in

column 1. Interestingly, the other controls also remain significant in the joint model. For example, papers by authors with 4-5 recent publications have about 85 log points higher citations than those with 0 recent publications. Interpreted through the lens of our model, this implies that referees evaluate papers as if they were substantially discounting the citations received by more prolific authors. There is a similar effect for papers with more co-authors and papers in more-cited fields.

**Mechanical Publication Bias**

So far, we have neglected the potential for a mechanical publication bias: papers that receive an R&R may accumulate more citations, conditional on quality, because the publication itself increases visibility, or provides a signal. This bias could lead us to overstate the impact of the determinants of citations. For example, positive referee recommendations may be correlated with citations not (only) because referees capture the paper quality, but because positive reports increase the probability that a paper obtains an R&R, which itself increases citations.

As we discussed in Section 2, under the assumptions of the model, we can address this issue with specification (7). We include an indicator for R&R, as well as a control function for the selection into the R&R stage, using as predictor the average R&R rate of an editor (with a leave-one-out mean). (This selection equation, which we discuss below, is in Column 9 of Table 2). The coefficient on the R&R indicator indicates the estimated mechanical publication effect (in log points), while the coefficient on the control function provides a measure of the "value added" of the editor.

We display the estimate in Column 4 of Table 2. The estimate on the control function is statistically significant and positive at 0.32 (s.e. 0.08). Through the lenses of our model, given an estimated $\sigma_v \approx 1.3$, this indicates a correlation of the editor signal with the paper quality of around 0.2. The estimated mechanical publication effect (the coefficient on the R&R dummy) of 0.06 (s.e. 0.14) indicates that the mechanical effect of an R&R is to increase citations by just 6 log points, an effect that is not statistically significant (if somewhat imprecisely estimated). If this effect seems too low, we stress that some of the papers receiving an R&R in our sample have not been published yet by the time the citations are collected in mid 2015, and that not all journals in our sample should be expected to have a sizable publication effect. We return to this point shortly.

Importantly, under this benchmark specification, the coefficients on the other variables—the referee recommendations, the author prominence, the number of authors — are barely affected compared to a specification without controls for R&R status and the selection effect (Column 3). On its face, this indicates that the bias arising from a mechanical publication effect is likely small.

A reasonable objection is that the specification above relies on a set of modeling assumptions. Can we provide a non-parametric upper bound for the effect of the mechanical publication bias? In Column 5, the specification includes the R&R control, but does not include the control function. In this specification, the higher citations for an R&R are entirely attributed to a mechanical publication bias, with no role for the editor value added. Under this upper bound, the mechanical publication bias is estimated to be 57 log points. Even under this extreme bounding assumption, importantly, the coefficients on the key variables are affected only to a small degree: the estimated effect of referee recommendations is about 20 percent lower, while the estimated effect of the prominence variables

is only 2-3 percent lower. Thus, even if we assume that the *entire* difference in citations for papers with an R&R is due to a mechanical effect, it does not explain much of the informativeness of referee reports, and of the other variables such as author prominence.

Column 6 presents the lower bound, assuming no publication bias but including the control function. The estimate is nearly identical to the estimate in Column 3 without the control function.

In Table 3 we present additional evidence expanding on the benchmark specification (Column 4 of Table 2), reproduced in Column 1 of Table 3. To the extent that there is a mechanical publication effect, we expect it to be larger for papers submitted earlier in the sample, since these papers have indeed gained in visibility from publication in a prominent journal. In comparison, given the typical length of the publication process, a paper submitted from 2011 onward that receives an R&R is unlikely to have been published for long by mid 2015 (when citations are collected), given the delays in the publication process. We also expect that the publication effect would be larger the higher the impact of the journal. By contrast, we would not necessarily expect differences in the editor value added along these dimensions.

Column 2 in Table 3 displays the evidence. The mechanical publication effect, is indeed, 49 log points larger for paper in the sample submitted up to 2010, as opposed to in the years from 2011 on. Further, the mechanical publication effect is larger for the highest-impact journal, the QJE, than in the other journals. Column 3 shows that there are no such interactions with respect to the control function term, suggesting that the value added is likely constant across these two dimensions.

The table reports also the coefficient on two representative determinants of citations, the fraction of R&R recommendations and an indicator for high prominence. These variables are essentially unaffected by the introduction of this more flexible model of the mechanical publication bias, compared to the benchmark model in Column 1.

We take these results as evidence of the presence of mechanical publication bias. On average, though, this estimated bias is not particularly large, and thus does not affect much the conclusions. Further, as we documented above, even under the upper bound, the coefficients on the determinants of citation are not much affected. Thus, in the rest of the paper we adopt as benchmark the specification with both publication bias and control function (Column 4 in Table 2).

**The Revise and Resubmit Decision**

Having examined the predictors of citations, we turn to the predictors of the R&R decision. As discussed in Section 2, under the joint assumptions that editors only care about the expected quality of papers and that citations are an unbiased measure of quality, the coefficients in a probit model for the R&R decision should be proportional to the coefficients in an OLS model for citations that includes the same variables. Under more general assumptions, however, this proportionality prediction will break down.

We first present some graphical evidence. Figure 4a (which is constructed like Figure 3a using paper×referee observations) shows that the probability of an R&R is strongly increasing in the recommendation of any one referee. To examine how editors aggregate multiple recommendations, we show a heat map in Figure 4b of the probability of an R&R verdict for all 49 possible combinations

of the referee recommendations when there are 2 referees. This probability is essentially zero with two negative recommendation, rises to 25 percent with two *Weak R&R* recommendations, and to 80 percent or higher with two *R&R* recommendations. Similar patterns are present looking at all possible pairs of recommendations for papers with three referees (Online Appendix Figure 3a). Along similar lines, Welch (2014) compares referee recommendations and editorial decisions for an anonymous journal and shows that editorial decisions are highly related to the referees' opinions.

Columns 7-9 of Table 2 present the estimated coefficients for probit models that parallel the citation models, using only the referee recommendations (column 7), only the other controls (column 8), and finally both sets of variables and the editor leave-out-mean R&R rate (column 9). As might be expected given the patterns in Figure 4a, the model with only the referee recommendations and journal×submission year controls is remarkably successful, with a pseudo $R^2$ of 0.48.[24] The quality of fit is apparent in the comparison between Figure 4c which plots predicted probabilities for each of the possible referee combinations for 2-referee papers, and Figure 4b, which shows the actual probabilities. The relatively close fit of the model across the cells is also true when we look at pairs of reports for papers with 3 referees (see Online Appendix Figures 3a-b).

Column 8 presents a model with only the $x_1$ (paper characteristic) variables. The R&R rate is increasing with the number of previous publications of the author team, but does not appear to be systematically affected by the number of coauthors, despite the effect of these variables on citations. The same is true of the field variables. Specifically, a comparison of the field effects in the R&R model and the citation model (reported in columns 1 and 3 of Online Appendix Table 2) shows little relation between the relative citations received by papers in a field and the relative likelihood the paper receives an R&R decision.

Column 9 presents the full specification of equation (4) with both the referee variables and the other covariates. This specification also includes the editor leave-out-mean average R&R rate. The addition of the latter variables raises the pseudo-$R^2$ of the probit very slightly (from 0.48 to 0.49), with most of the extra explanatory power coming from the author publication variables, which continue to exert a positive effect on the R&R rate, even controlling for the referee's recommendations. As in column 8, the number of authors has no systematic effect. The leave-out-mean variable has a statistically significant effect (t=2.9): there is variation among the editors in their R&R rate.

**Coefficient Plots**

With these results in hand, we turn to an examination of the relative magnitude of the coefficients of the various paper characteristics in the citation model and the R&R decision model. Our focus is on evaluating prediction P1, which states that if the editor is maximizing citations, the coefficients in these two models will be strictly proportional.

Figures 5a-b plot the coefficients from the R&R probit model (Column 9 of Table 2) against the corresponding coefficients from the citation model (column 4 of Table 2). For visual clarity, Figure 5a displays only the coefficients on the referee recommendation variables and on the author prominence variables, while Figure 5b shows all the coefficients. For interpretive purposes, the figures

[24]The journal-year fixed effects contribute very little to the fit, with a pseudo-$R^2$ of 0.03 when they are the only controls.

also show the best-fitting lines through the origin for various subgroups of coefficients. Under the null hypothesis of the model, these lines should all approximately have the same slope.

The referee recommendation coefficients in Figure 5a are remarkably aligned: referee categories that are associated with higher citations are also associated with a higher probability of an R&R decision. For example, the large jump in citations in moving from *Weak Revise and Resubmit* to *Revise and Resubmit* is mirrored by a large rise in the probability of R&R, while the negligible impact of moving from *Strong R&R* to an *Accept* recommendation on citations is also reflected by negligible effect on the probability of R&R. From this pattern one might conclude that the decision-making of editors is closely aligned to the views of the referees, and both are focused on higher citations.

When it comes to the other paper characteristics, however, the parallelism between citations and the R&R decision breaks down. For example, measures of author publications have a much smaller effect on probability of R&R than would be expected given their impacts on citations. The red line displays the degree of proportionality between the author publication variables in the R&R model and the citation model. The slope is only about one fifth the slope of the black line which shows the degree of proportionality between the referee recommendation coefficients in the models.

This conclusion is confirmed by a close examination of the coefficients in columns 4 and 9 of Table 2. The coefficients of the referee variables are about twice as big in the R&R model as they are in the citation model, implying in the context of our model that the standard deviation of the latent error in the editor's decision model ($\sigma_v$) is about 0.5 (since $\pi_R = \beta_R/\sigma_v$). In contrast, the coefficients of the author publication variables are only about 40% as large in the R&R model as the citation model. The two ratios differ by a factor of about 5, as is visible in Figures 5a and 5b.

Figure 5b also displays the coefficients of two other groups of variables: those associated with the number of authors and those associated with field. Both sets of variables have a significant effect on citations, yet editors put essentially no weight on the number of authors, nor do the coefficients on the field fractions appear to line up with their effects on citations (compare the coefficients in columns 2 and 4 of Online Appendix Table 2). Evidently, editors are putting much greater weight on referee recommendations relative to other variables that are also predictive of citations.

Do these patterns differ by journal? Online Appendix Figure 4 shows that several key patterns are common. (The underlying coefficients are reported in Online Appendix Table 3). First, within each group of variables, the coefficients line up nicely on a line. Second, the line for referee recommendations is systematically steeper than for other variables, implying that editors give more weight to the referee recommendations than to any of the other variables in forming their R&R decisions. Third, at all journals the measures of author publications have a particularly large and systematic impact on citations, but a much smaller relative impact on the R&R decision. This gap is particularly notable at REStud and REStat, where the editors appear to assign *no weight* to any variable other than the referee recommendations. Interestingly, this is consistent with the REStud's stated mission of supporting young economists.[25] At these journals, the R&R models seem to suggest that the editors simply follow the referees, with no attempt to undo any biases that referees exhibit in evaluating papers from different types of authors or from different fields.

[25]From the mission statement online: "[The] objective [of the Review] is to encourage research in theoretical and applied economics, especially by young economists".

**Visual Evidence on R&R and Rejects**

As an additional piece of graphical evidence, in Figure 6 we plot the average citation rate for papers that receive an R&R and papers that are rejected. For each paper we predict the probability of a revise-and-resubmit decision using the specification in Column 9 of Table 2. We then sort papers into deciles by this predicted probability, splitting the top decile into two top groups, and plot mean citations for papers with a positive and negative decision. We also show the number of papers in each probability range with each decision.

As shown along the $x-axis$ of the figure, for papers in the bottom 5 deciles of predicted citations the probability of an R&R is near zero, reaching just 1% in the fifth decile. The probability is still only 18% in the $8^{th}$ decile, but increases sharply to 37% in the $9^{th}$ decile and equals 90% for papers in the top 5 percent of submissions. The vertical gap between the mean citations for R&R's and rejected papers is relatively large – on the order of 60-80 log points. This vertical gap, as we discussed above, captures the combination of the mechanical publication bias and the editor value added. Interestingly, the vertical gap between R&Rs and rejects is wider to the left, as predicted by the model for informed editors: the editor has to receive a very positive signal for papers with relatively low observable quality in order to reach a positive R&R decision. Online Appendix Figure 8 displays the same data as in Figure 6 along with the predicted fit from our model, showing that the model does a good job of capturing the patterns in Figure 6.

Another salient feature of Figure 6 is that even among papers that receive a positive R&R recommendation, expected citations are increasing in the strength of the observable predictors. For example, mean $asinh(citations)$ for R&Rs in the top group in the figure (the top 5% of predicted citations) is about 4.1, while the mean for those in the 7th group (the top 60-70% of predicted citations) is about 3.6 – a gap of 50 log points. Thus, the close calls where the editor appears to have made a positive decision despite only lukewarm enthusiasm from the referees (and no offsetting $x_1's$) yield lower average citations than cases where the referees are very positive (and the editor agrees). This is consistent with the model and stresses the informativeness of referee recommendations.

**Other Citation Measures**

A potential concern with the findings so far is that the results hinge on our use of the inverse hyperbolic sine transformation in modeling citations. To address this concern, in Table 4 we reestimate the citation model using alternative transformations. Column 1 shows our base specification, reproduced from column 4 of Table 2. Column 2 uses our percentile citation measure, which controls for differences in citations across journal-year cohorts flexibly by computing citation percentiles within cohorts. Column 3 is motivated by the hypothesis that editors focus exclusively on the probability that a paper becomes a "major hit". Specifically, we define a paper to be *top cited* if it is in the top $p$ percent of citations in a journal-year cohort, where $p$ is set to the R&R rate for that journal and year. We then estimate a probit model to predict the probability of being top cited. Taking this point further, In Column 4, we use an indicator for a paper in the top 2% of citations in a journal-year cohort, proxying for "superstar" papers. We also consider a specification in column 5 using $\log(1 + citations)$ as an alternative to the asinh specification. Finally, in column 7 we re-

estimate our citation model using SSCI citations. Since SSCI citations only accrue to published papers, we restrict the sample to submissions in the years from 2006 to 2010 to ensure enough time for publication. To check the robustness of our main specification to the choice of sample, column 6 shows a model for $asinh(GS\,citations)$ fit to the 2006-2010 sample, which is similar to the baseline model in column 1.[26]

The results are very consistent across the alternative citation measures, with coefficients that are nearly proportional across specifications. For example, the coefficients in column 3 have a correlation of 0.998 with the coefficients in column 1, implying that the same index of observed paper characteristics predicts both mean asinh of citations and the probability of being in the upper tail of citations. In all cases referee recommendations are strong predictors of the measure of citations, with coefficients that are roughly proportional to the coefficients in the R&R probit, but of different scales depending on the citation measure. All the models also indicate significant positive effects of the author publication variables on the measure of citations, with a relative magnitude about 50% as large as the effects of the referee variables. Since the author publication variables enter the R&R probit model with coefficients only about 10% as large as the referee variables, we conclude that editors under-weight author publications by a factor of about 5 in their R&R decision, regardless of whether editors are maximizing expected asinh GS citations (column 1), the expected percentile of GS citations (column 2), the probability of being in the right tail of GS citations (columns 3 and 4), or the expected asinh or percentile of SSCI citations (columns 7-8). The one notable difference across specifications is that the estimated impact of the mechanical publication effect is much larger for the SSCI citations, as they should be, given that in this case there is indeed an obvious mechanical confound. This is confirmation of the fact that the Google Scholar variable is a more appropriate measure of paper quality, at least in our context.

**Additional Measures of Author Publications**

In our baseline specification we measure author productivity by the number of articles published in 35 high-impact journals over the 5-year period prior to submission. To probe the robustness of our under-weighting conclusion we checked three additional measures of productivity. The first is the count of publications in the previous 5 years in top-5 economics journals (REStud, QJE, the *American Economic Review*, *Econometrica*, and the *Journal of Political Economy,* excluding the Papers and Proceedings of the AER). The second is the count of publications in our 35-journal sample in the 6 to 10 years prior to submission. The third is an indicator for the prominence of the authors' home institutions, which may proxy for the quality of their past work or their promise as scholars (in the case of young researchers).

Table 5 presents citation models and R&R probit models in which we augment our baseline models from Table 2 (reproduced in columns 1 and 4) with these additional measures. The specifications in columns 2 and 5 include indicators for the number of author publications in top 5 journals. Since top-5 publications are relatively infrequent, we censor our measure at 4 publications in the past 5 years. As is evident from the estimates in column 2, measures of previous top-5 publications are

[26] As shown in Online Appendix Table 5, when we re-estimate our baseline R&R probit model using data from 2006-2010 the estimates are very similar to those from the whole sample period.

important predictors of citations: a paper from an author team with 2 recent top-5 publications is associated with an extra 44 log points of citations, even conditional on all the other variables. They also strongly affect the R&R decision. Nevertheless, their effect on the R&R decision relative to the effect of the referee recommendation variables is much smaller than in the citation model, suggesting a significant under-weighting of top-5 publications by editors relative to a citation-maximizing benchmark (see Online Appendix Figure 5f).

We also report the estimated effects of publications in the 35 high-impact journals in the period 6-10 years before submission. Although papers from authors with more publications in this earlier time frame do not receive significantly more citations (controlling for their recent publications), earlier publications do have a small positive effect on the R&R decision. Moreover, controlling for earlier publications and recent top-5 publications, the effects of recent publications in the broader 35 journal sample are all small and insignificantly different from 0.

Finally, in columns 3 and 6 we report the impacts of a measure of institutional prominence for the author team at the time of submission, distinguishing between US institutions (coded into 3 groups), European institutions (coded into 2 groups) and institutions in the rest of the world (coded into 2 groups). We use the rankings in Ellison (2013) to classify US institutions, while for non-US institutions we use the 2014 QS World University Rankings for Economics.[27] Since we only collected institutional prominence variables for REStud and JEEA, the models in columns 3 and 6 are fit to the subsample of submissions at these two journals.[28]

The results in column 3 show that institutional prominence is an important predictor of citations, even conditional on a broad set of measures of the authors' publication record. For example, having at least one coauthor at a top-10 US economics department at the time of submission increases citations by 51 log points, while having a coauthor at an 11-20 ranked US institution increases citations by 43 log points. Institutional affiliations also affect the R&R decision (column 6), but as with other characteristics included in $x_1$ their relative impact on the R&R decision is much smaller than the relative impact of the referee variables (see Online Appendix Figure 5g).

A particularly interesting set of findings concern the effects of institutional affiliation in Europe. Conditional on the referee recommendations, having a co-author at a top-10 department in Europe increases citations by 35 log points, a large and highly significant effect. Yet this affiliation has no significant effect on the R&R decision. Since *REStud* and *JEEA* are based in Europe, and many of the editors are drawn from top-10 European departments, this extreme downweighting cannot be explained by a lack of information about the relative standing of different schools. It appears that these two journals are "leaving citations on the table" by implicitly raising the threshold for an R&R decision when the author is from a top European department.Desk Rejections

While our main focus is the R&R decision, in this section we present a brief discussion of the desk rejection decision, building on the framework suggested by our simple model. An empirical analysis of this stage is useful given that more than half of the submissions to many journals are

[27]The institutional prominence dummies for each paper are defined within region, so that the dummies for each region sum to at most one, and the sum of the institutional dummies ranges from 0 to 3. Similar to our measure of author publications, we take the top-ranked U.S. institution among coauthors when defining the U.S. institution dummies, and the top-ranked European institution when defining the European dummies.

[28]Estimates of the models in columns 2 and 5 for these two journals are very similar to the ones for the full sample.

desk rejected, and that the previous empirical literature has largely ignored desk rejections.[29]

Using the full sample of 29,868 submitted papers, we compare predictors of citations with predictors of the decision to not desk reject (NDR) the paper. Author publications and the size of the author team are important predictors of citations (column 1 of Online Appendix Table 7). As would be expected if the NDR process selects papers based in part on the editor's private information about potential citations, the impacts of these variables are *larger* than when we estimate the same specification using only the subset of papers assigned to referees. For example, the coefficients of the publication measures in column 1 of Online Appendix Table 7 are approximately 1.3 times larger than the coefficients in the model in column 4 of Table 2, while the coefficients of the team size variables are about 1.1 times larger. These two sets of variables, plus field dummies and journal×year fixed effects have a combined R-squared of about 0.23 in predicting GS citations. Thus, there is considerable information in observed paper characteristics that can be used to predict citations.

A probit model for NDR, reported in columns 3-4 of Online Appendix Table 7, show that editors use the prior publication record of authors in making their initial NDR decision, but put little systematic weight on the number of co-authors or the field of the papers. A plot of the coefficients from the NDR probit against those of the citation model therefore shows systematic deviations from null hypothesis of citation maximization (Online Appendix Figure 7), with editors downweighting information in the number of coauthors and field relative to the information in prior publications.

**Value Added of the Editor at the Desk Reject Stage**

How much information does the editor have at the desk-rejection stage? This is a potentially important question because the desk rejection process is sometimes characterized as arbitrary or uninformed. Figure 7a plots mean citations for four groups of papers in various quantiles of the predicted probability of NDR. We show mean $asinh(citations)$ for papers that are desk rejected (the red line at the bottom) and those that are not desk rejected (the blue line) as well as separate lines for NDR paper that are ultimately rejected at the R&R stage (the green line) and those that receive a positive R&R decision (the orange line at the top of the figure).

The figure reveals large gaps in mean citations between desk-rejected and NDR papers, and between papers that are not desk rejected and then receive a positive or negative R&R.[30] On average, NDR papers receive about 75 log points more citations than those that are desk rejected, implying that the editor obtains substantial information from scrutinizing a paper before making the desk reject decision. In the context of our model this gap implies that the correlation between the editor's initial signal $s_0$ and future citations is about 0.32, and that $s_0$ reveals about 10% of the unexplained variance of citations given the observed characteristics at the desk reject stage.[31]

[29] On the theoretical side, Vranceanu et al. (2011) present a model in which papers with a poor match to the editorial mission of the journal are desk-rejected, but quality per se is irrelevant. Bayar and Chemmanur (2013) present a model in which the editor sees a signal of quality, desk rejects the lowest-signal papers, desk accepts the highest-signal papers, and sends the intermediate cases to referees. Schulte and Felgenhauer (2015) present a model in which an editor can acquire a signal before consulting the referees or not.

[30] The gap between papers that are R&R'd and those that are rejected after review is larger than the corresponding gap in Figure 6 (for the same set of papers) because of the different ways of grouping papers along the x-axis – by probability of NDR in Figure 7a (based only on $x_1$) and by probability of R&R in Figure 6 (based on $x_1$ and $x_R$).

[31] Recall that according to our model the signal to total variance ratio is $A_0 = \rho_0^2$, where $\rho_1 = 0.31$ is the implied correlation of the editor's signal and the citation residual.

The gap between NDR papers that are ultimately given an R&R and those that are rejected is also large – around 125-175 log points. This gap reflects the discriminatory power of the entire second stage of the review process, including the inputs of the referees and the editor's private signal at the R&R stage. For example, comparing papers that are reviewed by the referees and had an 80% probability of NDR based on $x_1$, those that ultimately receive R&Rs have mean $asinh(citations)$ of 4.0 while those that are ultimately rejected have a mean of 2.25 - implying about 5.7 times more citations for the R&R group.

Finally, the gap in average citations between desk rejected papers and those that are NDR but ultimately rejected is about 60 log points. This gap is interesting because both sets of papers are rejected - thus, there is no mechanical publication effect biasing the comparison. This gap can be decomposed as the sum of 75 log point gap in citations attributable to the NDR decision, minus a 15 log point gap attributable to the "bad news" of a rejection in the second stage. Viewed this way, the editor's signal at the desk reject stage is relatively informative.

So far, we have seen that author publications are highly predictive of the desk rejection decision. Since we do not have referee recommendations to benchmark the relative effect of the publication record, however, it is not clear whether editors over-weight or under-weight authors' publications in reaching their decision. Building on the test proposed by equation (12), we evaluate the hypothesis that desk rejection decisions are consistent with citation maximization by comparing citations for NDR papers with similar probabilities of desk rejection from more and less prolific authors.

We present this comparison in Figure 7b, focusing on authors (or author teams) with 4 or more recent publications versus those with 0 or 1 publications. Mean citations are about 100 log points higher for papers by more prolific authors, conditioning on NDR status *and* the quantile of the predicted probability of NDR. Indeed in most quantile bins the mean citations of desk rejected papers by more prolific authors have higher mean citations than the non-desk-rejected papers by less prolific authors. This pattern parallels our results at the R&R stage, where editors significantly under-weight the citations of more highly published authors, effectively imposing a higher bar (in terms of expected citations) for these authors. At both stages there appears to be a higher bar for authors with a stronger publication record.

# 5 Interpretation and Survey

To summarize our findings so far: at all four journals in our sample, referees and editors appear to impose a higher bar for papers by more prolific authors (or groups of authors). Figure 8a revisits the evidence for referees. We display mean $asinh(citations)$ for papers by more and less prolific authors with a given referee recommendation (using the same classification of prolific as in Figure 7b). If the referees were evaluating papers based on expected citations the two lines would be similar. Instead, mean citations for prolific authors are 100 log points higher. In other words, referees evaluate papers as if the citations received by more prolific authors should be discounted by roughly $e^1$.

Columns 1 and 2 of Table 6 quantify this discounting effect for the full set of publication dummies in our models. We begin in column 1 with a model for $asinh(citations)$, fit to the subsample of papers that were assigned to at least two referees, that includes only the author publication variables

and journal×year dummies. Relative to the omitted group of authors with no recent publications, papers from authors with 8+ publications have 139 log points higher citations. When we add in the referee recommendations and controls for field and number of coauthors (column 2), this gap falls by about 35 log points, but is still highly significant.

In the R&R decision model in column 9 of Table 2 we saw that editors put positive weight on author's publications (given the referee opinions), effectively "undoing" some of the bias against more prolific authors. How much does the R&R selection process reduce the effects of the publication variables? The answer is shown by the models in columns 3 and 4 of Table 6, which are fit to the subsample of papers that receive a positive R&R decision. Editors undo only about 20 percent of the implicit discounting imposed by referees, reducing the expected citation premium for papers from authors with 8+ previous publications, for example, from 105 to 82 log points. Interestingly, the estimated citation premiums for more prolific authors in the subsample of R&R papers are not very sensitive to whether we include the referee recommendations or not (as columns 3 and 4 show).

The estimated publication coefficients from the model in column 4 are very similar to the coefficients obtained when we implement the test described in Section 2.1, based on comparisons of citations for papers with the same probability of obtaining an R&R verdict. Specifically, we estimated a model for $asinh(citations)$ with dummies for papers in each decile of the predicted probability of an R&R verdict as well as an additional dummy for papers in top vingtile, and indicators for authors' previous publications. The estimated coefficient for 6 or more publications in this specification is 0.87 – quite close to the corresponding coefficient the model in column 4. (The other publication coefficients are also quite close). This confirms that we can clearly reject the hypothesis of citation-maximizing decision-making by editors.

To what extent does the citation advantage for papers of more prolific authors change when we condition on final publication status? While we do not know the publication status for all the R&R'd papers in our sample, we assume that the vast majority were ultimately published. We therefore used EconLit to construct a sample of all papers published in the 4 journals in our sample between 2008 and 2015. Assuming an average 2 year delay between first submission and publication, these papers should correspond to papers receiving an R&R in our sample from 2006 to 2013 (minus the papers that were rejected after an initial positive R&R verdict). We then constructed the $x_1$ variables for these papers, coding author publications at an assumed submission date 2 years before the publication date, and using the JEL codes in EconLit (which may differ from the codes at initial submission used in our main analysis). The estimated model using asinh GS citations as the dependent variable, shown in column 5 of Table 6, reveals a set of estimated publication coefficients that are slightly smaller than the ones in column 3 for R&R papers, but still large.[32]

Finally, for completeness we constructed a third sample of papers published in the top 5 economics journals between 1997 and 2012, coding the $x_1$ variables for these papers by assuming a 2 year lag between submission and publication, and using Google Scholar citations as of late 2016. Since these papers have all been published for at least 4 years, any concerns about publication-status-

[32]We measure Google Scholar citations in late 2016 for these papers using the same search protocols as for our main sample. We find 1,534 published papers in EconLit at the four journals, compared to 2,209 R&R recommendations. We believe the relative size of the published sample is reasonable, given that given that not all of the R&R papers are published and that the EconLit sample probably excludes most papers submitted to JEEA in 2003-05.

related biases are eliminated. Moreover, the model includes journal×year effects which control for differences in citations accruing to papers in more or less prestigious journals. The citation model for this sample (reported in column 6) yields estimated author publication effects that are attenuated by about 20-30% relative to the effects in our R&R sample (column 3), but are still highly significant.

## 5.1 Interpretations

Our model suggests two main interpretations for the key finding that referees and editors significantly under-weight the expected citations of papers by more prolific authors. The first is that citations are inflated measures of quality for prolific authors, leading referees and editors to discount citations accordingly. This could occur for several reasons. For one, more prolific scholars have broader networks of colleagues, students, etc., who know their work and cite it rather than a similar paper by some less prolific scholar. A closely related idea – Merton's (1968) "Matthew effect" – is that people tend to cite the best known author when there are several possible alternatives. Another possibility is that more prolific authors have more access to working paper series that publicize their work, inflating their relative citations, especially in the first few years after papers are written.

An alternative interpretation is that citations may be an appropriate measure of quality, but referees and editors impose a higher bar for more prolific authors. Such a process may be due to a desire to keep the door open to less established scholars (i.e., affirmative action) or a desire to prevent established authors from publishing marginally acceptable papers (i.e., animus).[33] We note that there are at least two pieces of evidence in the literature that support this interpretation. The most direct evidence is Blank's (1991) analysis of blind versus non-blind refereeing at the *American Economic Review,* which showed that blind refereeing increased the relative acceptance rate of papers from authors at top-5 schools. A second finding is that published papers written by authors who were professionally connected to the editor at the time of submission tend to have more rather than less citations (Laband and Piette, 1994; Medoff, 2003; Brogaard, Engelberg and Parsons, 2014).

Before we turn to some survey-based evidence designed to distinguish between these two interpretations, however, we briefly discuss a third possibility that is sometimes raised in the editorial context: elite favoritism.[34] According to this hypothesis, more accomplished authors are *favored* in the publication process by other prolific authors who review their work positively, and by editors who are in the same professional networks. If one takes citations as unbiased measures of quality, we clearly find substantial evidence against this hypothesis. It is possible, however, that the citations received by more prolific authors are highly inflated, and that after appropriate discounting (e.g., a discount of >100 log points) more prolific authors actually face a lower bar in the editorial process.

A plausible test of the elite favoritism hypothesis is to examine whether papers by prolific authors are evaluated more positively by other prolific scholars. Though all the editors in our sample have strong publication records, placing them squarely in the prolific category, the prior publication records of the referees vary widely. We thus test whether the citation gap in Figure 8a differs

[33] A related possibility is that editors may believe that less prolific authors are more likely to deliver a responsive revision if invited to provide one.

[34] This hypothesis is often raised informally by commentators who are skeptical of the integrity of the peer review process. See Campanario (1998a, 1998b) and Lee et al. (2013) for some context.

when the referee has a strong publication record (and is therefore a potential member of the elite) or not.[35] The comparison, shown in Figure 8b, gives no evidence of elite favoritism: the gap in citations between papers of prominent and non-prominent authors is about the same whether the recommendation comes from a prolific referee or a non-prolific referee. Interestingly, a similar conclusion was reached in the seminal study by Zuckerman and Merton (1971), which showed similar assessments of papers by more and less prominent authors by more and less prominent referees.

## 5.2 Survey Evidence on Quality vs. Citations

Using information on citations and the R&R decision we cannot distinguish between the two main explanations for the down-weighting of citations for papers written by more prolific authors. The two alternatives can be distinguished by data that allow us to measure quality independently of citations. In this section we present evidence from a survey designed with this purpose in mind.

The survey aims to replicate the quality assessment of referees and editors, using pairs of published papers. Specifically, the survey respondents compare two published papers that differ by the publication record of the author(s) at the time of submission, but are otherwise matched in terms of journal quality, publication year, and field. This contrast is designed to mirror the R&R decision faced by a journal editor in selecting among submissions. It also mirrors our empirical specifications which include controls for broad fields of the paper, as well as fixed effects for journal-year cohorts. The comparison of papers *within* a field also makes the evaluation easier for the survey respondents, and resembles the evaluation of referees who typically assess submissions in their field.

To identify pairs of papers, we consider articles published from one of the traditional top-5 journals in economics between 1999 and 2012, excluding AER Papers and Proceedings articles, notes, and comments. We code articles in the subfields of (i) unemployment; (ii) taxation; (iii) crime; (iv) education; (v) family economics; and (vi) behavioral economics.[36] We also code the articles as (mainly) theoretical or empirical.

Following the same procedure as in our main analysis sample, we measure the publications of authors in the same set of 35 high-impact journals in the 5 years prior to submission. Given the delays between submission and publication, we assume that papers were submitted 2 years prior to the year of publication. We then take, as in our main analysis, the maximum across all coauthors. We classify an author or author team as prolific if there is at least one coauthor with 4 or more publications in the 5 years prior to the assumed submission date. Likewise, we classify the author or team as non-prolific if none of the co-authors have more than 1 publication during this period. Notice that some of the authors coded as non-prolific at the assumed submission year may be coded as prolific in later years. This is as intended and reflects the procedure we used in our main analysis and the information available to the referees and editors at the time of submission.

We then identify balanced pairs of papers – one written by a prolific author, one by a non-prolific authors – published in one of the top-5 journals[37] in the same year, in the same field, and with the

[35] Berk, Harvey and Hirshleifer (2017) argue on the basis of interviews with former editors that relatively junior scholars are often harsher in all their reviews.

[36] The coding of the fields uses a combination of keywords. We search for the keywords in either the title of the paper, or in the description for one of the JEL codes associated with the paper.

[37] In constructing potential pairs we focused on papers from the *American Economics Review*, the *Quarterly Journal*

same theory or empirical component. To simplify our design we exclude papers by authors with intermediate publication records. We also exclude pairs with citations that were too imbalanced (a ratio of citations outside the interval from 0.2 to 5.0), and a small number of pairs that included a paper written by one of us, or that we viewed as too far apart in content. The final sample includes 60 pairs of papers, with 8 pairs on the topic of unemployment, 12 pairs on taxation, 6 pairs on crime, 12 pairs on education, 10 pairs on family economics, and 12 pairs on behavioral economics. The number of distinct papers is 101, since some papers appear as part of two pairs.

**Survey Wording.** The survey was administered on the Qualtrics platform, with all the questions displayed on one page (see Online Appendix Figure 9). The respondents were asked two main questions about each pair of papers they are asked to consider. The first asks their "opinion in comparing various features of the two papers," focusing on four specific criteria: (i) Rigor (theoretical structure and/or research design); (ii) Importance of Contribution; (iii) Novelty; (iv) Exposition (organization, clarity, detail, writing). For each criterion the respondent is asked to indicate whether Paper A is better, Paper A is slightly better, the two papers are about the same, Paper B is slightly better, or Paper B is better. We randomize the order in which the four criteria are asked, as well as whether Paper A or Paper B is the paper written by a prolific author.

Second, the survey informs the respondent of the Google Scholar citations as of August 2016 for the two papers and asks: *In light of the ____ citations accrued by Paper A and your assessment above, please indicate whether you think that the number of citations for Paper B is (i) about right, (ii) too high, (iii) too low.* We then elicit a quantitative measure of the appropriateness of citations:

> *In light of the ___ citations accrued by Paper A and your assessment above, what do you think the appropriate number of citations for Paper B should be?*

Let $c_A$ and $c_B$ denote the actual citations of papers A and B, and let $\hat{c}_B$ denote the elicited *appropriate number of citations* for paper B. When paper B is the one written by a prolific author, the ratio $\hat{c}_B/c_B$ represents the respondent's desired discount factor for the citations of the more prolific author. A value for this ratio that is less than 1 means that the respondent thinks the paper is "over-cited" relative to paper A, whereas a value greater than 1 means that he or she believes paper B is "under-cited". In the alternative case when paper A is the one written by a more prolific author, the desired discounting factor for citations to the paper by the more prolific author is $c_B/\hat{c}_B$.

The second half of the survey presents the same questions for a second pair of papers, and ends with an opportunity for the respondents to provide feedback.

**Survey Respondents.** The survey population includes faculty and PhD students who specialize in the fields covered by the papers in the survey. The survey was administered in September and October 2016. Our analysis follows a pre-registered analysis plan, AEARCTR-0001669.

Out of 93 emails sent to 73 faculty and 20 PhD students, 74 surveys were completed, 55 by faculty and 19 by PhD students, for an overall response rate of 80 percent. Each respondent compared 2 pairs of papers in their field, yielding $74 \times 2 = 148$ comparisons covering 58 distinct pairs.

*of Economics*, and the *Journal of Political Economy*, which tend to publish articles that are similar in the level of mathematical formality. For behavioral economics, given the smaller sample of articles, we include one article from *Econometrica*.

**Estimating the Mean Discount for Citations of More Prolific Authors**

For paper pair $j$, let $R_j$ represent the ratio of the number of citations for the paper written by the prolific author to the number of citations for the paper written by the non-prolific author. Using the respondent's answer to the question about the appropriate number of citations to paper B, we construct the respondent's quality-adjusted citation ratio as:

$$\begin{aligned}\widehat{R}_j &= \widehat{c}_B/c_A \text{ if paper B is by the prolific author} \\ &= c_A/\widehat{c}_B \text{ if paper A is by the prolific author.}\end{aligned}$$

We interpret $\widehat{R}_j$ as the respondent's assessment of the ratio of the quality of the paper by the prolific author in the $j^{th}$ pair ($q_{Pj}$) to the the quality of the paper by the non-prolific author ($q_{Nj}$), i.e.,

$$\widehat{R}_j = q_{Pj}/q_{Nj}.$$

Our model asserts that the relation between citations and quality is $\log c_{ij} = \log q_{ij} + \eta_{ij}$, where $i \in \{P, N\}$ and $\eta_{ij}$ reflects non-quality-related determinants of citations for paper $i$ in pair $j$. We assume that the within-pair gap in $\eta_{ij}$ can be decomposed as

$$\eta_{Pj} - \eta_{Nj} \quad = \quad \eta_\Delta + e_j$$

where $\eta_\Delta$represents average excess (log) citations accruing to papers by more prolific authors and $e_j$ is a random factor. It follows that

$$\log \widehat{R}_j = \log R_j - \eta_\Delta - e_j \tag{13}$$

Thus, we fit the simple regression model:

$$\log \widehat{R}_j = d_0 + d_1 \log R_j + \varepsilon_j. \tag{14}$$

According to our model we should estimate $d_0 = -\eta_\Delta$ and $d_1 = 1$.

A slightly more general model of the relationship between citations and quality is $\log c_{ij} = \theta(\log q_{ij} + \eta_{ij})$, which allows a concave or convex mapping from quality to citations. It is straightforward to show that all the implications of the model developed in Section 2 remain unchanged when $\theta \neq 1$.[38] In this case, however, equation (13) becomes:

$$\log \widehat{R}_j = \frac{1}{\theta} \log R_j - \eta_\Delta - e_j$$

and the predicted value for the coefficient $d_1$ in equation (14) is $d_1 = 1/\theta$.

Figure 9a illustrates two possible patterns of results using simulated data. We bin papers into

[38]The only change is that the coefficients in the citation model, equation (6), take on the values $\lambda_0 = \theta(\beta_0 + \eta)$, $\lambda_1 = \theta(\beta_1 + \eta_1)$, $\lambda_R = \theta\beta_R$, and the residual in the citation model becomes $\theta(\phi_q + \phi_\eta)$. Under citation maximizing behavior the coefficients of the R&R probit are still proportional to the coefficients in the citation model, but the factor of proportionality is $1/\theta\sigma_v$.

deciles by the citation variable ($\log R_j$) and plot the average of the y variable ($\log \widehat{R}_j$) within each bin. The dotted red line illustrates a case with, to a first approximation, no quality discounting: the regression line runs through the origin. The continuous blue line shows simulated data, assuming that papers by more prolific authors get 28 log points more citations than those by less prolific authors, on average, implying an intercept for the regression of $d_0 = -0.28$.

Figure 9b shows a bin-scatter of our actual data. Following our pre-analysis plan we winsorize the dependent variable at the $2^{nd}$ and $98^{th}$ percentiles. The average quality-adjusted citation ratios are clearly correlated with the actual citation ratios, with a slope close to 0.7 and an estimated intercept close to 0. Panel A of Table 7 provides a series of estimates of the model specified by equation (14), with a simple OLS regression in Column 1 and a specification in Column 2 in which we weight the responses for a given paper pair by the inverse of the number of respondents who evaluated the pair, thus giving equal weight to pairs evaluated by different numbers of respondents. In Column 3 we limit the sample to pairs with more comparable citations ($-0.5 \leq \log R_j \leq 0.5$). These three specifications suggest that holding constant quality, papers by more prolific authors receive between 1 and 3 percent more citations than those of less prolific authors.

In columns 4 and 5 we fit separate models for respondents who are either graduate students and younger faculty with relatively few publications (column 4) or faculty who would be classified as prolific (i.e., have published 4 or more papers in the past 5 years in one of the 35 journals in Online Appendix Table 1). The results show that any tendency to attribute excess citations to more prolific authors comes from prolific faculty, rather than from graduate students or faculty respondents with relatively few publications. This pattern provides no evidence of elite favoritism and suggests instead that the downweighting of citations to papers by relatively prolific authors may stem from competitiveness among prolific authors.

**Qualitative Ratings.** For paper pair $j$, the survey respondents also assess the relative strength of the two papers on a five-point scale, which we re-scale from -2 to +2 so positive values correspond to a higher rating for the paper by the prolific author. As shown in Figure 9c, there is at best a weak relationship between the respondents' assessments of the relative strengths of the papers and their relative citations $\log R_j$, with the the strongest relationship for relative importance (plotted with red dots). None of the scatters suggest a negative intercept, as would be expected if citations for more prolific authors are upward biased relative to quality.

Panel B of Table 7 presents regressions in which we relate the relative strength of the paper by the prolific author in a pair to the relative citation measure. Consistent with Figure 9c, only the model for "Importance" (column 2) has an R-squared above 0.05. Again, the key coefficient for our purposes is the constant, which (with a sign change) we interpret as an estimate of the excess citations received by more prolific authors, holding constant the relative quality in the particular domain. None of the estimated constants are large or even marginally significant, which is consistent with the result in Panel A. Overall, these results provide no evidence that papers by prolific authors receive more citations than those by non-prolific authors, controlling for their relative quality.

# 6 Conclusion

Editors' decisions over which papers to publish have major impacts on the direction of research in a field and on the careers of researchers. Yet little is known about how editors combine the information from peer reviews and their own prior information to decide which papers are published. In this paper, we provide systematic evidence using data on all submissions over an 8-year period for 4 high-impact journals in economics. We analyze recommendations by referees and the decisions by editors, benchmarking them against a simple model in which editors maximize the expected quality of the papers they publish, and citations are an ex-post measure of quality.

We show that this simple model model is consistent with some of the key features of the editorial decision process, including the systematic relationship between referee assessments, future citations, and the probability of an R&R decision, and the fact that R&R papers receive higher citations than those that are rejected, conditional on the referees' recommendations.

Nevertheless, there are important deviations from this benchmark. On the referee side, certain paper characteristics are strongly correlated with future citations, controlling for the referee assessments of a paper. This suggests that referees impose higher standards on certain types of papers, or that that they are effectively discounting the future citations that will be received by these papers. At best the editors only partially offset these tendencies. In particular, referees appear to substantially discount the future citations that will be received by more prolific authors and the editors offset the referees only slightly. Thus, among the papers that receive a revise-and-resubmit decision, those written by more prolific authors receive many times more citations, on average, than those written by less prolific authors, controlling for the referee assessment.

We consider two main interpretations. Citations may be an inflated measure of paper quality for prolific authors, leading referees and editors to discount citations accordingly. Alternatively, citations may be an appropriate measure of quality, but referees and editors may be using affirmative action to support less prolific authors. While our main analysis cannot separate the two interpretations, the results from a survey of economists asked to evaluate the quality of pairs of papers are most consistent with the affirmative action interpretation.

We view this just as a step in the direction of understanding the functioning of scientific journals, with many questions remaining. For example, are there similar patterns of citation discounting in other disciplines? Okike, Hug, and Kocher (2016) provide some evidence from a medical journal of favoritism towards prolific authors, a finding different from ours. Also, can a simple model explain the decision of editors to wait for another report or decide with what is at hand? We hope that future research will get at these and other questions.

# References


Aizer, Anna, and Joseph J. Doyle, Jr. 2015. "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges." *Quarterly Journal of Economics,* 130 (2): 759-803.

Bayar, Onur, and Thomas J. Chemmanur. 2013. "A Model of the Editorial Process in Scientific Journals." Working Paper.

Berk, Jonathan B., Campbell R. Harvey, and David Hirshleifer. 2017. "How to Write an Effective Referee Report and Improve the Scientific Review Process." *Journal of Economic Perspectives*, 31(1): 231-244.

Blank, Rebecca M. 1991. "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review." *American Economic Review*, 81(5): 1041-1067.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. Forthcoming. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy.*

Brogaard, Jonathan, Joseph Engelberg, and Christopher Parsons. 2014. "Networks and Productivity: Causal Evidence from Editor Rotations." *Journal of Financial Economics*, 111(1): 251-270.

Campanario, Juan Miguel. 1998a. "Peer Review for Journals as It Stands Today – Part 1." *Science Communications*, 19(3): 181-211.

Campanario, Juan Miguel. 1998b. "Peer Review for Journals as It Stands Today – Part 2." *Science Communications*, 19(4): 277-306.

Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature*, 51(1): 144-161.

Cherkashin, Ivan, Svetlana Demidova, Susumu Imai, and Kala Krishna. 2009. "The inside scoop: Acceptance and rejection at the journal of international economics." *Journal of International Economics*, 77(1): 120-132.

Chetty, Raj, Emmanuel Saez, and Laszlo Sandor. 2014. "What Policies Motivate Pro-Social Behavior? An Experiment with Referees at the Journal of Public Economics." *Journal of Economic Perspectives*, 28(3), 169-188.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad. 2014. "Family Welfare Cultures." *Quarterly Journal of Economics*, 129 (4): 1711-1752.

Ellison, Glenn. 2002a. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy*, 110(5): 947-993.

Ellison, Glenn. 2002b. "Evolving Standards for Academic Publishing: A q-r Theory." *Journal of Political Economy*, 110(5): 994-1034.

Ellison, Glenn. 2012. "Assessing Computer Scientists Using Citation Data." Working Paper.

Ellison, Glenn. 2013. "How Does the Market Use Citation Data? The Hirsch Index in Economics." *American Economic Journal: Applied Economics*, 5(3): 63-90.

Fisman, Raymond, Jing Shi, Yongxiang Wang, and Rong Xu. Forthcoming. "Social Ties and Favoritism in Chinese Science", *Journal of Political Economy*.

Griffith, Rachel, Narayana Kocherlakota, and Aviv Nevo. 2009. "Review of the Review: A Comparison of the Review of Economic Studies with its Peers." Unpublished Working Paper.

Hamermesh, Daniel S., George E. Johnson, and Burton A. Weisbrod. 1982. "Scholarship Citations and Salaries: Economic Rewards in Economics." *Southern Economic Journal*, 49(2): 472-481.

Hamermesh, Daniel S. 1994. "Facts and Myths about Refereeing." *Journal of Economic Perspectives*, 8(1): 153-163.

Heckman, James J. 1976. "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models." *Annals of Economic and Social Measurement*, 5(4), 475-492.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1), 153-161.

Heckman, James J., and Richard Robb, Jr. 1985. "Alternative methods for evaluating the impact of interventions: An overview." *Journal of Econometrics*, 30(1): 239-267.

Hilmer, Michael J., Michael R. Ransom, and Christiana E. Hilmer. 2015. "Fame and the fortune of academic economists: How the market rewards influential research in economics." *Southern Economic Journal*, 82(2): 430-452.

Hofmeister, Robert, and Matthias Krapf. 2011. "How Do Editors Select Papers, and How Good are They at Doing It?" *The B.E. Journal of Economic Analysis & Policy*, 11(1): article 64.

Knowles, John, Nicola Persico, and Petra Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy*, 109(1): 203-229.

Laband, David N., and Michael J. Piette. 1994. "Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors." *Journal of Political Economy*, 102(1): 194-203.

Larivière, Vincent, Véronique Kiermer, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A simple proposal for the publications of journal citation distributions." *bioRxiv* preprint.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in Peer Review." *Journal of the American Society for Information Science and Technology*, 64(1): 2-17.

Li, Danielle. 2017. "Expertise vs. Bias in Evaluation: Evidence from the NIH" *American Economic Journal: Applied Economics*, 9(2): 60-92.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review*, 103(5): 1797-1829.

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior". In *Frontiers in Econometrics*, edited by Paul Zarembka, 105-142. New York: Academic Press.

Medoff, Marshall H. 2003. "Editorial Favoritism in Economics?" *Southern Economic Journal*, 70(2): 425-434.

Medoff, Marshall H. 2006. "Evidence of a Harvard and Chicago Matthew Effect." *Journal of Economic Methodology*, 13(4): 485-506.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science*, 159(3810): 56-63.

Okike, Kanu, Kevin T. Hug, Mininder S. Kocher, and Seth S. Leopold. 2016. "Single-blind vs Double-blind Peer Review in the Setting of Author Prestige." *JAMA: The Journal of the American Medical Association*, 316(12): 1315-1316.

Schulte, Elisabeth, and Mike Felgenhauer. 2015. "Preselection and Expert Advice." Macie Paper Series Working Paper.

Seglen, Per O. 1997. "Why the impact factor of journals should not be used for evaluating research." *BMJ: British Medical Journal*, 314(7079): 498-502.

Smart, Scott, and Joel Waldfogel. 1996. "A citation-based test for discrimination at economics and finance journals." NBER Working Paper 5460.

Vranceanu, Radu, Damien Besancenot, and Kim Huynh. 2011. "Desk rejection in an academic publication market model with matching frictions." ESSEC Working Paper.

Welch, Ivo. 2014. "Referee recommendations." *Review of Financial Studies*, 27(9): 2773-2804.

Wooldridge, Jeffrey M. 2015. "Control Function Methods in Applied Econometrics." *Journal of Human Resources*, 50(2): 420-445.

Zuckerman, Harriet, and Robert K. Merton. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System." *Minerva*, 9(1): 66-100.