What do Editors Maximize? Evidence from Four Economics Journals*

David Card
UC Berkeley and NBER

Stefano DellaVigna
UC Berkeley and NBER

August 2018

Abstract

We study editorial decisions using anonymized submissions at 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 and referees 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 these recommendations quite closely. We document two main deviations from a citation-maximizing benchmark. First, papers by highly-published authors get more citations, conditional on the referees' recommendations, suggesting that referees set a higher bar for these authors or that prolific authors are over-cited conditional on quality. Editors' decisions at the desk-reject and revise-and-resubmit stage show a similar pattern. Second, recommendations of highly-published referees are no more predictive of future citations, yet editors give their views significantly more weight. To help interpret these findings we collect two additional pieces of evidence. We conduct a survey asking field specialists to assess the relative quality of matched papers by more and less prolific authors; our respondents do not appear to believe that prolific authors are over-cited. We also elicit forecasts of the informativeness of referees from a sample of editors and other economists. Both groups believe that highly-published referees are more informative, potentially explaining the extra weight they receive in editors' decisions.

*We thank Daron Acemoglu, Pierre Azoulay, Esther Duflo, Glenn Ellison, Joey Engelberg, Patricia Funk, Joshua Gans, Matthew Gentzkow, Daniel Hamermesh, Campbell Harvey, David Hirshleifer, Nagore Iriberri, Lawrence Katz, Chris Parsons, Imran Rasul, Laszlo Sandor, Jesse Shapiro, Scott Stern, Vera te Velde, Ivo Welch, and Fabrizio Zilibotti for comments and suggestions. We thank Luisa Cefala', Alden Cheng, Bryan Chu, Jared Grogan, Johannes Hermle, Kaushik Krishnan, Patricia Sun, Andrew Tai, and Brian Wheaton for outstanding research assistance. We also acknowledge the generous support of the editors and staff at the four journals in our database, and thank respondents to our survey. Our survey was approved by UC Berkeley IRB, protocol 2016-08-9029 and pre-registered as trial AEARCTR-0001669.

1 Introduction

Editorial decisions at top academic journals help shape the careers of young researchers and 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 referees and editors set the same bar 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 and referees, whether the paper was desk rejected or sent to referees, summary recommendations of the referees, and the editor's reject or revise-and-resubmit decision. All submissions, regardless of the editor's decision, are matched to citations from Google Scholar and the Social Science Citation Index.

These unique data allow us to significantly advance our understanding of the decision process at scientific journals. Most previous research has focused on published papers or aggregated submissions (e.g., Laband and Piette 1994; Ellison 2002; Hofmeister and Krapf, 2011; Card and DellaVigna, 2013; Brogaard, Engelberg, and Parsons, 2014). While these studies offer many insights, they cannot directly illuminate 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 studies (Cherkashin et al. 2009 and Griffith, Kocherlakota, and Nevo 2009) present broader analyses for two journals, 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 the referees' recommendations, the characteristics of a paper and its authors, and their own private information to determine which papers to invite for revision. As a starting point we assume that editors maximize the expected quality of published papers and that quality is revealed by citations – i.e., citation-maximizing behavior. While this benchmark has some appeal, given the salience of impact factors to editors and publishers and the importance of citations for promotions and salaries, it has at least three major limitations.[1] First, the mere publication of a paper in a prestigious journal may raise its citations, introducing a mechanical publication bias. Second, editors may set higher or lower thresholds for certain groups of authors.[2] Third, even in

[1] See Seglen (1997) and Larivière et al. (2016) for a discussion of impact factors, and Ellison (2012), Hamermesh, Johnson, and Weisbrod (1982) and Hilmer, Ransom, and Hilmer (2015) for analysis of how citations affect promotions and salaries in economics.

[2] Laband and Piette (1994), Medoff (2003), and Brogaard, Engelberg, and Parsons (2014) all find that submissions to economics journals 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.

the absence of editorial preferences, citations may be systematically biased by differences in citing practices across fields, or by a tendency to cite well-known authors (Merton, 1968).

We incorporate all three features in our modeling framework and econometric specifications. First, we allow for a direct impact of the R&R decision on ultimate citations. Using differences in R&R rates across editors at the same journal (analogous to the *judges design* used in many recent studies) we develop a control function to separate the mechanical effect of R&R status 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 our key results. Second, we allow referees and editors to hold preferences for or against certain types of papers, relaxing the "citation maximizing" objective. Finally, we allow for the possibility that papers by certain authors (or in certain fields) may receive more citations, holding quality constant.

We focus our main analysis on the R&R decision for non-desk-rejected papers. Papers at the journals in our sample are typically reviewed by 2 to 4 referees who provide summary evaluations ranging from *Definitely Reject* to *Accept*. Consistent with a citation-maximizing benchmark, the referee recommendations are strongly predictive 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 find that editors' R&R decisions are heavily influenced by the referees' recommendations: the summary 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 with each summary recommendation 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 editors largely follow the referees, papers invited for revision have significantly higher citations conditional on the referee reports and other characteristics, suggesting that editors have private information about the quality of papers. Within our model, this implies a correlation of the editor's private quality signal with the unobserved determinants of citations of 0.20.

Nevertheless, there are two important deviations from citation maximization. First, the referee recommendations are *not* sufficient statistics for expected citations, even within field. In particular, submissions from prolific authors receive substantially more citations, controlling for referee recommendations. For example, papers by authors with 6 or more recent publications (in a set of general interest and field journals) have on average 100 log points more citations than papers with similar referee ratings by authors with no recent publications. This gap is essentially unchanged when we use a control function to adjust for any mechanical publication bias, and is only slightly smaller under an extreme bound. This suggests that referees impose higher standards on papers by prolific authors, or that they effectively discount the future citations that will be received by these papers.

Editors' R&R decisions reveal that they value papers by prolific authors more than the referees, but only slightly so: at all four journals we find that editors undo at most one-quarter of the penalty imposed by the referees. We conclude that editors either agree with the referees that there should be a higher bar for more prolific authors, or they agree with the referees that papers by these authors get too many citations, conditional on quality.

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

The second key deviation regards the weight that editors place on the recommendations of different referees. Measuring informativeness by the strength of the correlation between recommendations and citations, we find that referees with 3 or more recent publications are *equally informative* as referees with fewer publications. Nevertheless, editors place put significantly more weight on the recommendations of referees with more publications. This finding is similar when we allow the informativeness of referees to vary by journal and by field of the paper and is not explained by how many reports the referees have done for the editor.

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 little is known about how DR decisions are made. Editors appear to have substantial private information at the DR stage: conditional on field, author publication record, and other factors, papers sent for refereeing accumulate many more citations than the papers that are desk rejected. Even papers that end up rejected after refereeing have 72 log points more citations on average than papers that are desk rejected. Since both groups of papers are ultimately rejected, this comparison bypasses any concern about endogenous publication effects. As at the R&R stage, editors also appear to discount the expected citations for papers by more prolific authors. Conditional on the probability of desk-rejection, desk-rejected papers by prolific authors have higher average citations than non-desk-rejected papers by authors with no previous publications.

In the final part of the paper, we return to the interpretation of the two key deviations from the benchmark model. Our finding that referees *and* editors act as if they under-value citations to papers by more prolific authors runs counter to some earlier research and to the common perception that the publication process is if anything biased in favor of prominent scholars.[4] In an attempt to disentangle the two competing explanations for this finding we conducted a survey of faculty and PhD students in economics, asking them to compare matched pairs of papers in their field – one written by more prolific authors, the other written by authors who at the time of submission had few prior publications. We provide respondents with the actual GS citations for each paper and ask them to evaluate the appropriate citation ratio, given the quality of the papers. Interestingly, our respondents' preferred relative citations for prolific authors are only 2% below their actual relative citations (standard error = 5%). We interpret this as evidence that the observed penalty on expected citations at the R&R stage more likely reflects a *higher bar* for prolific authors than a discount for the fact that their papers get too many citations, conditional on quality.

We then turn to interpreting the second result, on the differential reliance of editors on prolific referees. One explanation for this finding rests on incorrect beliefs: editors expect the recommen-

[4]See Lee et al. (2013) for a recent review of the literature outside economics. In economics most previous work has found surprisingly weak evidence of such bias, or even evidence of a higher bar for prominent scholars. Blank (1991)'s 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 these authors when their identities were known. Smart and Waldfogel (1996) and Ellison (2011) 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 interpret as measures of editorial treatment. Similarly, 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. Medoff (2006) finds that papers by authors from Harvard and the 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.

dations of highly-published referees to be more informative and therefore rely on them more. An alternative is that editors want a *quiet life*: they find the recommendations equally informative, but are reluctant to ignore or overturn the recommendations of prolific referees.

In order to provide evidence on the incorrect beliefs interpretation, and more generally to assess how much is known about editorial decision-making, we collect forecasts as in DellaVigna and Pope (forthcoming) from a group of editors and associate editors at the *Review of Economic Studies* and a set of faculty and graduate students. Editors are aware that they set a higher bar for papers by prolific authors at the DR stage, a pattern that faculty do not anticipate. Editors, faculty, and students all believe that highly-published referees are better able to forecast citations than less published referees, providing support for the biased beliefs interpretation.

In light of these results, we believe our findings do provide insights into the editorial process – even for editors themselves – and lay the groundwork for a deeper understanding of this process, at least in the upper tier journals in economics.

2 Model

To help organize our empirical analysis we develop a stylized model of the editorial decisions, focusing mostly on the R&R stage. For simplicity we ignore any stages after the 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 referee recommendations x_R. Quality is determined by an additive model:

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

where the unobserved component (ϕ_q) is normally distributed with mean 0 and standard deviation σ_q. For the moment we ignore the possibility that the referee assessments may be more or less informative depending on characteristics of the referees or the paper. We also ignore the possibility that papers assigned to different types of referees may have higher or lower quality (e.g., Hamermesh, 1994, Bayar and Chemmanur, 2013). We come back to both issues in our empirical model.

Note that in specifying equation (1) we do *not* assume that the referee recommendations are unbiased forecasts of quality, conditional on the observable characteristics x_1. There are at least two possible explanations for the role of observable characteristics in equation (1). One is that the referees observe noisy signals of paper quality and simply report their signals to the editor, rather than Bayesian estimates of quality that incorporate any prior information contained in x_1. In this case $\beta_1 x_1$ can be interpreted as a scaled version of the prior mean for quality.[5] An alternative is that the referees believe that papers with different characteristics should meet different quality

[5]To see this, assume there is one referee who observes a noisy signal s^R, and that the prior mean for quality is $\psi_1 x_1$. The posterior mean of quality is $\lambda s^R + (1-\lambda)\psi_1 x_1$, where $\lambda \in (0,1)$ is a shrinkage factor that reflects the noise in s^R. If the referee reports her signal to the editor then $\beta_1 = (1-\lambda)\psi_1$. We thank Glenn Ellison for this point.

thresholds and adjust their recommendations accordingly. In this case $\beta_1 x_1$ measures the differences in the referees' quality thresholds for different papers.

The editor observes a signal s which is the sum of ϕ_q and a normally distributed noise term ζ with standard deviation σ_ζ:

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

Conditional on s and $x \equiv (x_1, x_R)$ the editor's forecast of ϕ_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 ϕ_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 reject the paper or not. 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 give a positive decision ($RR = 1$) for papers with $E[\log q | s, x] \geq \tau_0$, where τ_0 is a fixed threshold that depends on the target acceptance rate.[6] This acceptance rate τ_0 will depend on the journal and year, accounting for example for different R&R rates across the journals and years. An editor who accepts papers with the highest chance of exceeding a given quality threshold would follow the same rule.[7]

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 canonical 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:

$$P\left[RR = 1 | x\right] \quad = \quad \Phi\left[\frac{\beta_0 - \tau_0 + (\beta_1 - \tau_1)x_1 + \beta_R x_R}{\sigma_v}\right] \tag{4}$$

[6] 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.

[7] Since the editor's posterior is normal the expected probability of exceeding a quality threshold q^* is $\Phi\left[\frac{\beta_0 - log(q^*) + \beta_1 x_1 + \beta_R x_R}{\sigma_v}\right]$. Selecting papers for which this probability exceeds a certain bound leads to the same decision rule as choosing those for which $E[\log q | s, x]$ is above a certain threshold.

$$= \quad \Phi\left[\pi_0 + \pi_1 x_1 + \pi_R x_R\right],$$

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 η :[8]

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

The simplest assumption is that η depends only how long a paper has been circulating: in this case citations form a perfect index of quality apart from an adjustment for the age of the paper. More generally, 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 ϕ_η :

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

The coefficient η_{RR} measures any mechanical bias arising because R&R papers are likely to be published sooner (and in a higher ranked journal) than those that are rejected. Combining equations (5) and (1) leads to a simple 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 $\phi = \phi_q + \phi_\eta$.

When η is constant across papers (i.e., $\eta_1 = \eta_{RR} = 0$) we can recover β_1 and β_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 λ_1 in equation (6) will reflect both quality (q) and η. Moreover, OLS estimation of equation (6) poses a problem because RR status is endogenous and will be positively correlated with the error component ϕ to the extent that editors' private signals are informative about quality.

To recover consistent estimates of the coefficients λ_1, λ_R and λ_{RR}, we assume that different editors have different quality thresholds for reaching an R&R decision (i.e., different values of τ_0) but that the particular editor assigned to a paper has no effect on citations. In this case, following the judge assignment approach (e.g., Maestas, Mullen, and Strand, 2013; Dahl, Kostøl, and Mogstad, 2014), we can use the R&R rate for other papers handled by the same editor as a variable that shifts the threshold for R&R but has no independent effect on citations.

For our main specifications we augment equation (6) with a control function that represents the generalized residual from the editor's RR decision model (Heckman and Robb, Jr., 1985; Wooldridge, 2015). Specifically, we first fit a probit model for the R&R decision, including x_1, x_R and the

[8] As we discuss in the Online Appendix 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 $\theta = 1$.

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:

$$r = \frac{(RR - \Phi[\pi(x,z)])\,\phi[\pi(x,z)]}{\Phi[\pi(x,z)]\,(1-\Phi[\pi(x,z)])}$$

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 the instrumental variable z. Finally, we include $\widehat{r}$ (the estimate of r) in the citation model:

$$\log c = \lambda_0 + \lambda_1 x_1 + \lambda_R x_R + \lambda_{RR} RR + \lambda_r \widehat{r} + \phi'. \tag{7}$$

The inclusion of $\widehat{r}$ absorbs any endogeneity bias in RR status. Moreover, the estimate of λ_r provides a measure of the correlation $\rho_{v\phi}$ between the editor's private signal (v) and the unobserved determinants of citations (ϕ) 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 λ_r to estimate the informativeness of the editor's signal. Otherwise, the implied correlation will tend to under-estimate ρ_{vq} because citations contain an extra component of noise. In the online Appendix we present maximum likelihood estimates that estimate the R&R probit and the citation model jointly: these are very close to the to the two-step estimates.

Two concerns with this procedure are that the identity of the editor assigned to a paper may affect citations (controlling for journal and field) or that our functional form assumptions are incorrect. In either case the estimated coefficients for x_1 and x_R will be potentially biased. To address these concerns, we derive an upper bound and re-estimate equation (7) without including $\widehat{r}$, thereby ignoring the likely 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 benchmark model with two simplifying assumptions:

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

Under these assumptions a comparison of equations (4) and (7) shows that the editor's weights in the R&R decision rule will be *strictly proportional* to the weights in the citation model:

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

If we graph the estimates $(\widehat{\pi}_1, \widehat{\pi}_R)$ from the R&R probit against the corresponding estimates $(\widehat{\lambda}_1, \widehat{\lambda}_R)$ from the model for log citations, the points will lie on a positively sloped line that passes through the origin with slope $1/\sigma_v$.

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 attempts to uncover.

Regardless of the source of non-proportionality, we can quantify the net citation penalty or premium imposed by editors on papers with a given characteristic. Consider the k^{th} element of the vector x_1 and let π_{1k} represent the coefficient of this characteristic in the R&R probit model. Under a citation-maximizing benchmark, $\tau_{1k} = \eta_{1k} = 0$ and the coefficient of this characteristic in the citation model would be $\lambda_{1k} = \pi_{1k}\sigma_v$. More generally, the difference between the coefficient in the citation model and the expected coefficient under citation maximizing behavior is:

$$\lambda_{1k} - \pi_{1k}\sigma_v = \tau_{1k} + \eta_{1k} \equiv \theta_{1k}. \tag{8}$$

The term θ_{1k} measures the excess effect of the k^{th} characteristic on citations relative to the benchmark posed by the editor's decision rule. We refer to this gap as the editor's "citation penalty" for papers with characteristic x_{1k}. As discussed below, we can estimate the θ_{1k} coefficients for key paper characteristics (such as the authors' previous publication record) by jointly estimating the R&R probit and the citation model and imposing the proportionality assumption $\pi_R = \lambda_R/\sigma_v$ for the effect of the referee recommendations, yielding an estimate of σ_v.

2.2 The Desk Reject Decision

At the earlier desk-rejection stage the only observable information is x_1. In the Online Appendix we develop a simple model that closely parallels our model for the R&R decision illustrating how editorial preferences and/or systematic biases in citations as measures of quality determine the relative effects of x_1 on future citations and the probability of non-desk-rejection (NDR).

Under the citation-maximizing benchmark, expected citations should be *the same* for any two papers with the same probability of non-desk-rejection by a given editor. Specifically, expected citations, conditional on x_1 and non-desk-rejection ($NDR = 1$), should be a function only of the probability of NDR, $p(x_1)$:

$$E\left[\log c | x_1, NDR = 1\right] = G(p(x_1)), \tag{9}$$

where $G(\cdot)$ is a strictly increasing continuous function. Equation (9) leads to a simple 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

with different values of an individual covariate (such as the author's previous publication record).[9] Under citation maximization, $p(x_1)$ is a sufficient statistic for expected citations and there should be no difference in expected citations for papers in a cell. If, instead, editors are using a different threshold for different authors or citations are a biased measure of quality, 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 to assemble an anonymized data set of submissions with information on the year of submission, approximate field (based on JEL codes at submission), the number of coauthors and their recent publication records, the summary recommendations of each referee (if the paper was reviewed), the publication record of each referee, an (anonymized) identifier for the editor handling the paper,[10] 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.[11]

All four journals use the Editorial Express (EE) software system, which stores information in standardized files that can be accessed by a user with managing editor permissions. We developed a program that extracted information from the EE files, queried the GS system, and merged publication histories for each author and referee from a data base of publications in major journals (described below). The program was run on a stand-alone computer under the supervision of an editorial assistant and created 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.[12] We constructed our data sets for 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=4,946).

As Table 1 and Figure 1a show, 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).

[9] 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).

[10] Our access agreement with REStat did not allow us to retain editor identifiers, but we were allowed to retain an indicator for higher and lower-tenure editors, which we use to form two groups of editors. For the other journals we combine editors who handled very few papers.

[11] The data set does not include any information on demographic characteristics of the authors or referees, such as age, year of highest degree, or gender, and does not track authors or referees across papers.

[12] The top-code limit for citations is 200 for REStud and 500 in the other journals. We adjust for the lower top code at REStud using an imputation procedure based on the mean of citations at the other journals for papers that are above the REStud topcode. We also show below that accounting for the censoring does not change the results.

The R&R rate is lowest at the QJE (4%) and highest at REStat (12%).[13]

Figure 1b 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*.[14] The modal recommendation is *Reject* at all four journals. Between 54% (REStat) and 73%(QJE) or all recommendations are *Definitely Reject* or *Reject*.[15]

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.

To measure prior publications of authors, we built a database of all articles published between 1991 and 2014 in 35 high-quality journals (including the leading general interest journals and the top field journals for a majority of fields – see Online Appendix Table 1). We then merge authors to this data base and count the number of papers published in a 5-year window ending in each year from 1995 to 2013.[16] For papers with multiple authors we take the highest publication count among all coauthors, setting the count to 0 if we find no previous publications. We also record the number of coauthors, since this variable is highly correlated with citations (Card and DellaVigna, 2013).

As shown in Table 1 and Figure 1c, 46% of papers in our overall sample were submitted by author teams 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. We follow a similar procedure to assign publication records to referees. As Figure 1d shows, referees tend to have more publications than authors.

We recorded the number of citations received by a paper 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 as follows. 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 under the allintitle search, and verify that a given GS entry has at least one author surname in common with the authors in EE. Then the GS and SSCI citation

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

[14] We combine the top two categories, *Conditionally Accept* and *Accept*, which are rare, into the *Accept* category.

[15] Table 3 in Welch (2014) 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.

[16] We also calculate the number of 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.

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.[17]

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 rise over time. 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.[18] Online Appendix Figure 1 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).

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.[19]

Figure 2a shows how *asinh* of the number of citations is related to individual referee recommendations. 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. There is a strong positive 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 degree of referee informativeness across journals. The *levels* of the citation measure differ, however, with the highest citation levels at the QJE, reflecting differences in the submission pool, the desk-rejection rate, and the effectiveness of the desk-rejection decision.[20]

How do citations vary with the collective opinions of the entire team of referees? Online Appendix Figure 3a presents a heat map of mean citations for papers with 2 reports for each of the 7×7=49 possible cells for the two referee recommendations.[21] The figure reveals that average citations depend on the average opinions of the referees. For example, papers receiving two *Reject* recommendations

[17] Our main results are robust to dropping from the analysis papers with no match in GS.

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

[19] In particular, papers at REStud assigned to only one referee have a 99% rejection rate. We therefore exclude the 2,264 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.

[20] Online Appendix Figures 2a-c show similar patterns for alternative citation measures.

[21] 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.

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. We find similar evidence for papers with 3 reports (Online Appendix Figure 3c).

In light of this evidence, we summarize the recommendations using the fractions of recommendations for a given paper in each of the 7 categories. For example, if a paper 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. We generalize this approach below to allow for potential differences in the weight given to different referees.

Column 1 in Table 2 reports the estimates of an OLS regression model for $asinh(citations)$ that includes journal $\times$ year fixed effects and the fractions of reports in each category. The estimates show a strong correlation between average referee evaluations and mean citations, with larger coefficients than implied by the slopes in Figure 2a. This reflects the fact that the coefficients in the regression model measure the effect of having all referees unanimously select a given recommendation, whereas the figure measures the effect for only a single referee.

To document the validity of our averaging specification we return to the subsample of papers with two reports, and show in Online Appendix Figure 3b that the predicted citations from the model in Column 1 of Table 2 are very similar to the actual citations in Online Appendix Figure 3a. The model also does well for papers with 3 reports (Online Appendix Figures 3c-d). Moreover, as shown in Online Appendix Table 2, 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 determinants of citations: 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, a paper with 3 coauthors has 25 log points more citations. There are also systematic differences in citations acoss fields (see Online Appendix Table 3): 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 for published papers (e.g., Card and DellaVigna, 2013).

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 in Column 1. Importantly, the other controls also remain significant in the joint model, though smaller in magnitude than in the specification in Column 2. For example, papers by authors with 4-5 recent publications have about 85 log points higher citations than those with 0 recent publications, controlling for the recommendations, versus about 111 log points without controlling for x_R. There are similar effects for papers with more coauthors 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. 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.

Under the assumptions of the model, we can address this issue with specification (7). In Column 4 we include an indicator for R&R, as well as a control function for the selection into the R&R stage, using as an instrumental variable the leave-out mean R&R rate of the editor. (This selection equation, which we discuss below, is reported in Column 10 of Table 2). The coefficient on the R&R indicator gives the mechanical publication effect (in log points), while the coefficient on the control function provides a measure of the "value" of the editor's signal.

The estimated coefficient for the control function is positive and significant ($t \approx 4$). Given the residual variance of citations ($\sigma_\phi \approx 1.6$), the 0.32 point estimate implies a correlation of the unobservable determinants of the editor's decision with the unobserved component of citations of around 0.2. The coefficient on the R&R dummy of 0.07 (s.e.=0.14) indicates that the mechanical effect of an R&R is to increase citations by just 7 log points, though we cannot rule out an effect as large as 35 log points. In interpreting this estimate we stress that many of the papers receiving an R&R in our sample were not published by the time we collected citations in mid 2015, and that not all journals in our sample would be expected to have a sizable publication effect relative to alternative outlets. We return to these points shortly.

Importantly, the coefficients on the other variables—the referee recommendations, prior author publications, and the number of authors — are barely affected by the addition of the control function and R&R dummy. This suggests that any biases in these coefficients arising from mechanical publication effects are small.

A reasonable objection to this specification is that it relies on the assumption that a particular editor has no direct effect on citations. To probe the robustness of our conclusions we use the bounding approach described above and include the R&R dummy while excluding the control function. This yields an upper-bound estimate of $\eta_R = 0.57$ (Column 5 of Table 2). Even under this extreme assumption the coefficients on the other key variables are only modestly affected: the estimated coefficients of the referee recommendations are about 20 percent smaller than those in column 3, while the estimated coefficients of the author publication variables are 2-3 percent smaller.

Online Appendix Table 4 presents a series of additional checks on the publication effect. We expect a larger mechanical publication effect for submissions in the early years of our sample, since these papers had more time to benefit from publication, and also for higher-impact journals. As Column 2 shows, the mechanical publication effect is, indeed, 50 log points larger for earlier submissions than for later submissions, and is larger for the highest-impact journal, the QJE, than for the other journals. Reassuringly, the informativeness of the editor's signal is instead similar across journals and cohorts (Column 3). Importantly, the other coefficients in the model are highly robust.

We interpret these results as confirming that there are indeed positive effects of receiving an R&R on subsequent citations, particularly at the QJE and for papers reviewed earlier on. On average, though, this effect is not particularly large, and it does not impact our conclusions regarding the relative size of other determinants of citations, even under an upper-bound assumption. In the rest of the paper we therefore adopt as benchmark the simpler specification in Table 2.

The Revise and Resubmit Decision

We now turn to the predictors of the R&R decision. Figure 2b (which is constructed like Figure 2a using $paper \times referee$ observations) shows that the probability of an R&R decision is strongly increasing in the recommendation of any one referee. Notice that the probability of R&R only rises to 0.4 or 0.6 even for a *Strong R&R* or *Accept* recommendation because this relationship does not condition on the recommendations in other reports. This also explains why the relationship is flatter for the QJE where, on average, editors use a higher number of referees.

To examine how editors aggregate multiple recommendations, Online Appendix Figure 4a displays a heat map 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 *Reject* recommendations, rises to 25 percent with two *Weak R&R* recommendations, and to 80 percent or higher with two *Revise and Resubmit* recommendations. Along similar lines, Welch (2014) shows that referee recommendations and editorial decisions are highly correlated for an anonymous journal.

Columns 8-10 of Table 2 present the estimated coefficients for probit models that parallel the citation models, using only the referee recommendations (Column 8), only the x_1 variables (Column 9), and finally both sets of variables and the editor's leave-out-mean R&R rate (Column 10). As might be expected given the patterns in Figure 2b, the model with only the referee recommendations and $journal \times submission\ year$ controls is remarkably successful, with a pseudo R^2 of 0.48.[22] The quality of fit is apparent in the comparison between Online Appendix Figure 4b, which plots predicted probabilities for each of the possible referee combinations for 2-referee papers, and Online Appendix Figure 4a, which shows the actual probabilities.[23]

The specification in Column 9 shows that the R&R rate is increasing with the number of previous publications of the author team, but is not systematically affected by the number of coauthors, despite the positive impact of these variables on citations. Similarly for the field variables, a comparison of the coefficients in the R&R model and the citation model (Columns 1 and 3 of Online Appendix Table 3) 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 10 presents the full specification of equation (4). Relative to the model in Column 8 the full specification has a slightly higher pseudo R^2. Moreover, the author publication variables have a significant effect, as does the editor's leave-out-mean R&R rate ($t = 3.9$).[24]

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

[23] The close fit of the model across the cells is also evident when we look at pairs of reports for papers with 3 referees (see Online Appendix Figures 4c-d).

[24] As noted earlier, our confidentiality agreement at REStat precluded editor identifiers but allowed us to retain an indicator for editors in their 4th year or later of tenure. Within each year of submissions, we treat the two tenure groups as separate "editors".

Comparing Determinants of Citations and the R&R Probability

Figure 3a plots the coefficients from our baseline R&R model (Column 10 of Table 2) against the corresponding coefficients from the citation model (Column 4). For visual clarity, we display only the coefficients on the referee recommendation variables and the author publication variables, and also show the best-fitting lines through the origin for the two subgroups of coefficients. Under the hypothesis of citation maximization, these lines should have the same slope, equal to $1/\sigma_v$, the inverse standard error of the editor signal that is the unobserved component in the R&R model.

The referee recommendation coefficients in Figure 3a 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 editors closely follow the referees and that both are focused on higher citations.

When it comes to other paper characteristics, however, the parallelism between citations and the R&R decision breaks down. Measures of author publications have a much smaller effect on the probability of R&R than would be expected given their impacts on citations. The relative slope of the red line (summarizing the author publication coefficients) versus the blue line (summarizing the referee recommendations coefficients) is only about 0.20. In other words, comparing the relative effects of various factors in citations versus the R&R decision, author publications are downweighted by a factor of 5 in the R&R decision.

Interpreted in the context of our model, this suggests that editors only partly offset the tendency of referees to discount the expected citations of more prolific authors. For example, consider the effect of an author team with 4-5 recent publications. From the baseline citation model model (column 4 of Table 2) these papers receive 85 log points more citations, controlling for the referees' opinions. Using the framework of equation (8), the expected citation premium for these papers in the absence of any editorial preference or excess citation effects is $\pi_{1k}\sigma_v$, where $\pi_{1k} = 0.31$ is the estimated coefficient associated with these papers in our baseline R&R probit (column 10 of Table 2). Using the inverse of the slope of the line in Figure 3a for the referee recommendation variables we can estimate $\sigma_v \approx 3/7 = 0.43$. Thus, based on the premium editors place on papers by authors with 4-5 recent publications we would expect a citation premium of only $0.31 \times 0.43 = 0.13$, which is far smaller than the actual premium. Our estimate of editors' citation penalty is therefore $\theta_{1k} = \lambda_{1k} - \pi_{1k}\sigma_v = 0.85 - 0.13 = 0.72$ (72 log points), which is 85% as large as the penalty implicit in the referee recommendations. [25]

Do these patterns differ by journal? Online Appendix Figure 5 (based on the coefficients in Online Appendix Table 5) shows that several key patterns are common. Within each group of variables the coefficients for each journal fall nicely on a line. Also, the line for referee recommenda-

[25] We can perform this same calculation using the upper bound model in Column 5 of Table 2: the implied editor penalty is very similar (73 log points), illustrating the robustness of our conclusions to the treatment of publication bias. We can also estimate similar discount factors for other author groups. For example, the editor penalty for authors with 6+ publications is 82 log points, versus the 101 log point effect implicit in the referees' opinions.

tions is systematically steeper than for the author publication variables, implying that editors give more weight to the referee recommendations than to the author publication variables in forming their R&R decisions, for a given impact on citations. The disparities are particularly striking at REStud, where the editors appear to assign essentially *no weight* to variables other than the referee recommendations.[26] Interestingly, this lack of attention to prior publications is consistent with the REStud's explicit mission of supporting young economists.[27]

Visual Evidence on Citations for R&R and Rejected Papers

As an additional piece of graphical evidence, in Figure 4a we plot the average citation rate for papers that receive an R&R and for those that are rejected. For each paper we predict the probability of a revise-and-resubmit decision using the specification in Column 10 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.

As shown along the x axis of the figure, for papers in the bottom 5 deciles 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 90% in the top ventile. The vertical gap between the mean citations for R&R's and rejected papers is large: 60-80 log points. This gap captures the sum of the mechanical publication effect and the editor's value added (i.e., $\lambda_{RR} + \lambda_r \widehat{r}$). It is wider to the left of the figure, as predicted: the editor has to receive a very positive signal (leading to a large positive value for $\widehat{r}$) to reach a positive decision for papers with low observable quality. Online Appendix Figure 6 displays the same data along with the predicted fits from our model, and shows that the model does a good job of capturing the patterns in Figure 4.

Another salient feature of Figure 4a is that even among papers that receive an R&R, expected citations are increasing in the strength of the observable predictors. For example, mean $asinh(citations)$ for R&Rs in the top ventile are 50 log points higher than for R&Rs in the 7^{th} decile. In the context of our model this means that the editor's signal is only partially informative, so the "selection bias" implied by a positive R&R decision ($\lambda_r \widehat{r}$) is not large enough to compensate for low levels of the observable factors. It is also interesting that average citations for R&R'd papers in the 5th and 6th deciles are slightly above average citations for rejected papers in the top ventile – so the editorial decision process is broadly consistent with citation maximization.

Figure 4b breaks down the two groups of papers—R&Rs and rejected papers—by a measure of author publications and shows that rejected papers by more prolific authors have about the same citations as R&Rs by less prolific authors. In particular, rejected papers in the top ventile by more prolific authors out-perform R&Rs by less prolific authors in the 5th and 6th decile by about 50 log points, implying a sizable cost in terms of citations of the deviation from citation maximization.

[26] Again, using the framework of equation (8) we can quantify the degree of discounting applied by editors to papers by more prolific authors. The editor penalty for authors with 4-5 recent publications is about 70 log points overall, is about 65 log points at QJE, 90 log points at REStud, 74 log points at REStat, and 33 log points at JEEA.

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

Informativeness of Different Referees

So far we have assumed that different referees are all equally informative, and that editors assign them equal weights in making a decision. In the language of the model, the coefficients β_R and π_R do not depend on the characteristics of the referees. Yet it is plausible that referees who have themselves published more papers are better able to judge papers. It is also plausible that some types of papers are easier or harder to judge.

Figure 5a presents graphical evidence along the lines of Figure 2a on the informativeness of reports, distinguishing between recommendations from referees with 3 or more recent publications and referees with <3 recent publications. Average citations are monotonically increasing in the recommendation for both groups, with a very similar slope, suggesting that reports by more and less prolific referees are equally informative. The two lines differ only in their intercepts: for each level of referee enthusiasm the papers assigned to prolific referees have about 20 log points more citations, presumably because editors tend to assign more promising papers to prolific referees.

In contrast to the parallel lines in Figure 5a, Figure 5b shows that editors appear to place more weight on the recommendations of more prolific referees. Papers that receive a *Definitely Reject* or *Reject* recommendation by either group of referees are very unlikely to receive an R&R verdict, but more positive assessments by prolific referees appear to have greater impact than similar assessments by less prolific referees.

To proceed further, we move to a regression-based framework that allows us to control for other characteristics that differ between the papers assigned to more or less prolific referees. To keep the specification manageable, we assume that the summary recommendation of the j^{th}referee of a given paper (x_{Rj}) are scaled by common vector λ_R that does not vary across referees or papers, yielding a one-dimensional index $\lambda_R x_{Rj}$. We then allow this index to be up-weighted or down-weighted by a "slope factor" $exp(\xi z_j)$ that depends on a set of characteristics z_j of the referee and the paper (so z_j can include x_1). Letting J denote the number of referees assigned to a given paper and denoting $z = \sum_{j=1}^{J} z_j$, our extended citation model becomes:[28]

$$asinh(c) = \lambda_0 + \lambda_1 x_1 + \lambda_z z + \frac{1}{J}\sum_{j=1}^{J} exp(\xi z_j)\lambda_R x_{Rj} + \lambda_{RR} RR + \lambda_r \widehat{r} + \phi. \tag{10}$$

The model in Column 2 of Table 3 presents a simple version of this specification in which z_j includes an indicator for a referee having 3+ recent publications as well as a set of journal dummies. The estimated slope coefficient for prolific referees confirms the pattern in Figure 5a: referee assessments of more prolific referees are no more informative about citations than those from less-prolific referees, though papers assigned to a higher share of prolific referees have higher average citations. As shown in Column 3, these results are robust to adding a set of field dummies to z_j.[29]

In Columns 6-7 we present parallel specifications for the R&R probit (i.e., we amend equation (4) by including a term than weights the referee recommendations by a slope factor, as in equation (10)). Again confirming the visual impression from Figure 5b, we find that editors put about 20

[28]Notice that since $x_R = \sum_j x_{Rj}/J$, when $\xi = 0$ this amounts to our baseline model with the addition of controls for the mean value of z_j.

[29]As Online Appendix Table 6 shows, econometrics and theory have flatter citation-recommendation relationships.

percent more weight on the recommendations of prolific referees.[30] Figure 3b plots the coefficients from the R&R probit model in Column 7 against the coefficients from the citation model in Column 3, distinguishing between the relative effects of recommendations from more and less prolific referees. Since the recommendations of more prolific referees get more weight in editors' decisions but are no more informative, the coefficients for these referees (the lighter line) lie on a line that is about 20% steeper than the corresponding line for the less prolific referees (the darker line). This pattern holds in each of the four journals in the sample (Online Appendix Table 7).

A possible interpretation of the higher weight for prolific referees is that editors pay more attention to referees who do more refereeing for an editor, and these happen to be the more prolific economists. (This does not explain why they get more weight, but highlights an alternative channel). Thus, in Columns 4 and 8 we also control for *asinh* (Previous Reports), where "Previous Reports" is the number of reports a given referee has provided to the editor handling the paper (from the start of our sample period). Interestingly, the recommendations of referees with more previous reports tend to be weighted more by editors **and** forecast citations better, by about the same magnitude. Nevertheless, the addition of this control leaves the puzzling pattern for prolific referees essentially unaltered. We return below to the interpretation of this key deviation from citation maximization.

Robustness Checks: Alternative Measures of Citations and Author Publications

In this section we investigate the robustness of our main findings to two key measurement issues: how we measure citations; and how we assess authors' track records at the time of submission.

In two key checks, we consider the impact of highly-cited papers. First, one may be concerned about the censoring of citations at 500 (200 for REStud). In Column 6 of Table 2 we present a Tobit specification that explicitly models the censoring; this leaves the results nearly unaffected. Next, we take into account the fact that arguably editors are particularly interested in predictors of "superstar" papers, since such papers contribute disproportionately to the impact factor. A probit model predicting papers in the top 2% within a journal-year cell (Column 7 in Table 2) yields coefficient estimates that are parallel to the estimates from our baseline specification.

Online Appendix Table 8 presents estimates of additional alternatives to our baseline citation model from Column 4 of Table 2 using alternative dependent variables: (i) the *percentile rank* of GS citations for a paper within journal $\times$ submission-year groups; (ii) an indicator for *top cited* papers, equal to 1 if a paper is in the top p percent of citations in a journal-year cohort, where p is the R&R rate for that journal and year; (iii) an indicator for a paper in the top 5% of citations in a journal-year cohort as an alternative measure of "star" papers; (iv) $\log(1 + citations)$; (v) *asinh* of SSCI citations;[31] and (vi) *asinh* of GS citations, excluding papers for which we could not find any Google Scholar results (instead of assigning 0 cites to those papers). The results are consistent across the alternative measures, with coefficients for the referee recommendation and author publication

[30] The estimated $\hat{\xi}$ for the referee publication variable is significantly different in the citation regression and in the probit model, as we show with a bootstrap.

[31] Since SSCI citations only accrue to published papers, we restrict the sample to submissions in 2006-2010 to ensure enough time for publication. We checked that estimates from our basic model using *asinh* of GS citations are quite similar when we restrict the analysis to this same sample period.

record variables that are nearly proportional across specifications.[32]

Next, we consider three alternative measures of author productivity: (i) a count of publications in the 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); (ii) a count of publications in our 35-journal sample in the 6 to 10 years prior to submission; and (iii) indicators 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).

As Online Appendix Table 9 shows, previous top-5 publications are important predictors of citations, even conditional on all the other variables, and they also affect the R&R decision. Their effect on the R&R decision relative to the effect of the referee recommendation variables is much smaller than on citations, implying a significant under-weighting of top-5 publications by editors relative to a citation-maximizing benchmark. In contrast, publications in the 35 high-impact journals in the 6-10 years before submission have little or no value in predicting citations (controlling for recent publications), but they do have a small positive effect on the R&R decision. This is the only instance of an author publication variable that is *over*-valued by editors relative to its effect on citations.

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).[33] Institutional prominence is an important predictor of citations, even conditional on 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. Institutional affiliations also affect the R&R decision, 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.

An interesting set of findings concern the effects of institutional affiliation in Europe. Conditional on the referee recommendations and other controls, having a coauthor 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 downweighting cannot be explained by a lack of information about the relative standing of different schools.[34]

[32] One interesting difference is that the estimated effect of initial R&R status is large and positive in the model for SSCI citations (estimate = 0.77; standard error = 0.15), which makes sense since SSCI only records citations to published papers.

[33] 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. The institutional prominence dummies 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. Since we only collected institutional prominence variables for REStud and JEEA, these models are fit to the subsample of submissions at these two journals. Estimates of the models in Columns 3 and 6 for these two journals are very similar to the ones for the full sample.

[34] As a final robustness check, we ask whether the citation premium for papers from prolific authors appears also in a sample of published papers. Given that the editor undoes the premium only partially, we would expect this to be the case. Indeed, as Online Appendix Table 10 shows, there is a similar (with smaller magnitudes) citation premium for prolific authors for published papers in the four journals in our sample, and in top-5 journals.

Structural Estimates

Our main specifications in Table 2 are derived from a two-step procedure: we first estimate a probit model for the editor's R&R decision (with a single extra variable - the leave-out mean R&R rate of the editor) and then estimate an OLS model for asinh citations including the generalized residual from the probit and an indicator for R&R status. As noted, however, we can estimate the two equations jointly by maximum likelihood, imposing the key structural assumption of our model that the referee recommendation variables enter proportionately in the R&R and citation models (i.e., that $\lambda_R = \pi_R \sigma_v$). This allows us to directly estimate the editor's citation penalty factors specified by equation (8) (i.e., the θ_{1k} coefficients), although we cannot separately identify the contributions of editor preferences versus excess citations relative to paper quality.

The results from this structural estimation are summarized in Online Appendix Table 11. We show the estimated coefficients in the editor's R&R model (the π coefficients), as well as the implied estimate of σ_v and the estimated citation penalties, focusing on three key sets of variables: the referee recommendation variables, the author publication variables, and the indicators for the number of authors of the paper. (The model also includes unrestricted journal×year and field dummies in both equations). We also show the implied coefficients of the citation model (i.e. the λ coefficients, where $\lambda_{1k} = \pi_{1k}\sigma_v + \theta_{1k}$). For comparison, the table also presents the coefficients of our baseline R&R decision model and citation model.

The implied coefficients from the structural model are very similar to those from the unrestricted 2-step model. This reflects the fact that, as shown in Figure 3a, the key structural assumption of proportionality in the effects of the referee recommendation variables in the R&R probit and the citation model is approximately true in the data. Indeed, the structural estimate of σ_v is 0.39 (standard error=0.03) which is quite close to the inverse slope of the best fittting line joining the referee recommendation variables in Figure 3a.

Perhaps the most interesting feature of the structural model is how large are the editor discount factors relative to the implied discount factors in the referee assessments of different papers. For example, the editor discount applied to citations of authors with 4-5 publications is 73 log points (essentially the same as the estimate derived informally from the unrestricted estimates in Table 2), compared to the 86 log point discount implied by the referee assessments. For the variables representing the number of authors, the editor and referee discounts are effectively the same. We conclude that referees and editors both tend to undervalue papers by authors with more prior publications, or by larger teams of co-authors, relative to the citations these papers will receive.

4.2 Desk Rejections

While our main focus is the R&R decision, in this section we present a brief discussion of the desk rejection (DR) decision, building on the simple model described in the Online Appendix and summarized in Section 2.2. An empirical analysis of DR's is useful given that more than half of the submissions to many journals are desk rejected, and that the previous empirical literature has

largely ignored desk rejections.[35]

Using the full sample of 29,872 submitted papers, we compare predictors of citations with predictors of the decision to not desk reject (NDR) the paper in Online Appendix Table 12. Author publications and the size of the author team are important predictors of citations (Columns 1-3). Editors use the prior publication record of authors in making their initial NDR decision (Column 4), but put little systematic weight on the number of coauthors or the field of the papers. These estimates provide additional evidence of deviations from the 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.

How much information does the editor have at the desk-rejection stage? This is an important question because the desk rejection process is sometimes characterized as arbitrary or uninformed. Figure 6a plots mean $asinh(citations)$ for four groups of papers in various quantiles of the predicted probability of NDR: (i) papers that are desk rejected (the red line at the bottom), (ii) papers that are not desk rejected (the blue line), (iii) NDR papers that are ultimately rejected at the R&R stage (the green line), and (iv) NDR papers that receive an R&R (the orange line at the top).

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, conditional on the estimated probability of NDR.[36] On average, NDR papers receive about 80 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.[37]

The gap in average citations between desk rejected papers and those that are NDR but ultimately rejected is 72 log points. This gap is interesting because both sets of papers are rejected - thus, there is no mechanical publication effect biasing the comparison. 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 unclear whether editors over-weight or under-weight authors' publications in reaching their decision. Building on the test proposed by equation (9), 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.

[35] On the theoretical side, Vranceanu, Besancenot, and Huynh (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. In Bayar and Chemmanur (2013)'s model, 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. In Schulte and Felgenhauer (2015)'s model, an editor can acquire a signal before consulting the referees or not. Our simple model can be interpreted this way.

[36] The gap between papers that are R&R'd and those that are rejected after review is larger than the corresponding gap in Figure 4 (for the same set of papers) because of the different ways of grouping papers along the x-axis – by probability of NDR in Figure 6a (based only on x_1) and by probability of R&R in Figure 4 (based on x_1 and x_R).

[37] As shown in the Online Appendix, in our NDR model the signal-to-total-variance ratio of the editor's signal before making a desk reject decision is $A_0 = \rho_0^2$, where $\rho_0 = 0.32$ is the correlation of the editor's signal and the citation residual. Thus $A_0 \approx 0.10$.

We present this comparison in Figure 6b, focusing on authors (or author teams) with 3 or more recent publications versus those with 0-2 publications. 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. At both stages there appears to be a higher citation bar for authors with a stronger publication record.

5 Interpretation and Additional Evidence

In this section we return to the two key deviations from citation maximization. First, referees and editors appear to impose a higher citation bar for papers by prolific authors. Second, recommendations by prolific referees are equally predictive of the citations of a paper, but editors put more weight on the recommendations of prolific referees. We discuss potential interpretations for these findings and provide additional evidence from two surveys of economists designed to help distinguish between the alternative interpretations.

5.1 Citation Bar for Prolific Authors

There are two main explanations for our first key finding that referees and editors significantly under-weight the expected citations of papers by more prolific authors. The first is that papers by prolific authors are *over-cited*, leading referees and editors to discount their citations accordingly. Over-citing could arise because more prolific authors have better access to working paper series and other distribution channels that publicize their work, inflating their citations. They may also have networks of colleagues and students who cite their work gratuitously, or cite it instead of similar work by less prolific scholars. Finally, people may tend to cite the best known author when there are several possible alternatives - Merton (1968)'s "Matthew effect."

An alternative interpretation is that citations are unbiased measures of quality, but referees and editors set 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 marginal papers (i.e., animus).[38] At least two pieces of evidence in the literature support this interpretation. Blank (1991)'s analysis of blind versus non-blind refereeing at the *American Economic Review* showed that blind refereeing increased the relative acceptance rate of papers from authors at top-5 schools. Second, published papers written by authors who were professionally connected to the editor at the time of submission tend to have more rather than fewer citations (Laband and Piette, 1994; Medoff, 2003; Brogaard, Engelberg, and Parsons, 2014).

A third hypothesis, elite favoritism, holds that 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.[39] If one takes citations as unbiased measures of quality, we find substantial evidence against this hypothesis. It is possible, however, that the citations received

[38] A related possibility is that editors impose a higher bar for prolific authors because they believe these authors will be less willing to revise their paper to accommodate the referees' and editors comments.

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

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.

To gather evidence on this hypothesis, we examined whether papers by prolific authors are evaluated more positively by other prolific scholars. Online Appendix Figure 8 does not support elite favoritism: the gap in citations between papers of prominent and non-prominent authors is the same whether the recommendation comes from a prolific referee (a possible member of the elite) or from a non-prolific referee.[40]

5.1.1 Survey Evidence on Quality vs. Citations

To attempt to distinguish between the two leading explanations, we conducted a survey designed to measure quality separately from citations. We asked economists to compare matched pairs of papers in the same topic area, published in the same year in a similarly-ranked journal. The comparison was designed to mirror the R&R decision faced by a journal editor in selecting among submissions. It also mirrors the design of our main empirical models, which include controls for field and fixed effects for journal-year cohorts. The key difference is that our survey respondents were asked to evaluate the relative quality of papers, *not* to make R&R recommendations. Thus, we hoped to abstract from any tendency to raise or lower the bar for prolific authors at the refereeing stage. We describe the key design choices and results, with additional information in the Online Appendix.

We selected paired sets of papers from articles published in a top-5 journal between 1999 and 2012 in 6 topical areas. Following the same procedure as in our main analysis, we measure the publications of authors in the 35 high-impact journals in the 5 years prior to submission, assuming that papers were submitted 2 years prior to the year of publication. We classify authors/author-teams as prolific if at least one coauthor has 4 or more publications in the 5-year period, and as non-prolific if none of the coauthors have more than 1 publication during this period. We then selected balanced pairs of papers – one written by a prolific author, one by a non-prolific author – published in one of the top-5 journals in the same year, in the same field, and with the same relative mix of theoretical versus empirical content. We exclude potential 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 other pairs, as detailed in the Online Appendix. The final sample included 60 pairs of papers.

We sent the survey (Online Appendix Figure 9) to faculty and PhD students in the relevant fields in the Fall of 2016. Our analysis followed 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.

The respondents were asked two main questions about each pair of papers. First they compared features of the two papers, such as rigor, importance, and novelty. Second, they made a quality judgment as follows. The survey informs the respondent of the GS citations as of August 2016 for the two papers[41] and asks: "*In light of the ___ citations accrued by Paper A and your assessment*

[40] Interestingly, the seminal study by Zuckerman and Merton (1971) also found that more and less prominent referees tended to give similar assessments of papers by more and less prominent authors.

[41] We debated whether to provide the citations numbers. In the end, we decided to do so to provide respondents

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 for ease of exposition consider the case in which paper B is the one written by a prolific author (the order was randomized). For paper pair j, $R_j = c_B/c_A$ is 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, $\widehat{c}_B$, we construct the ratio $\widehat{R}_j = \widehat{c}_B/c_A$. We interpret $\widehat{R}_j$ as the respondent's assessment of the relative quality of paper written by the prolific author in pair j, that is, $\widehat{R}_j = q_{Pj}/q_{Nj}$.

Our model assumes the citation-quality relation $\log c_{ij} = \log q_{ij} + \eta_{ij}$, where $i \in \{P, N\}$. We decompose the within-pair gap in η_{ij} as $\eta_{Pj} - \eta_{Nj} = \eta_\Delta + e_j$, where η_Δ represents average excess (log) citations accruing to papers by more prolific authors and e_j is a random factor. It follows that

$$\log \widehat{R}_j = \log R_j - \eta_\Delta - e_j \tag{11}$$

Thus, we fit the simple regression model:

$$\log \widehat{R}_j = d_0 + d_1 \log R_j + \varepsilon_j. \tag{12}$$

According to our model we should estimate $d_0 = -\eta_\Delta$, and thus the intercept provides a measure of the quality discount for citations for prolific authors, measured in log points.[42]

Figure 7 displays a bin scatter plot of the elicited quality ratio ($\log \widehat{R}_j$) against the actual citation variable ($\log R_j$). The two variables are clearly correlated, with a slope close to 0.7 and an estimated intercept—our measure of the average degree of quality discounting for prolific authors—very close to 0. Panel A of Online Appendix Table 13 provides estimates of the model in equation (12), with an OLS regression in Column 1 and a specification in Column 2 in which we weight the responses by the inverse of the number of respondents who evaluated the pair. In Column 3 we limit the sample to pairs with more comparable citations ($-0.5 \leq \log R_j \leq 0.5$). Holding constant quality, papers by more prolific authors receive 1-3 percent more citations than those of less prolific authors.

In Columns 4 and 5 we fit separate models for graduate students and younger faculty with relatively few publications (Column 4) or faculty who would be classified as prolific. Interestingly, any tendency to attribute excess citations to more prolific authors comes from prolific faculty, rather than from graduate students or faculty respondents with few publications. This pattern suggests a potential role for competitiveness among prolific authors in explaining the results.

We similarly find no evidence of quality discounting for papers by prolific authors using the qualitative ratings (Panel B of Online Appendix Table 13). Overall, these results provide little evidence that papers by prolific authors are over-cited, controlling for relative quality.

with all the relevant information, which they could have easily obtained in any case with a quick search.

[42]The model also makes the prediction that $d_1 = 1$. In the Online Appendix we show that a slight generalization of our citation model, with $\log c_{ij} = \theta(\log q_{ij} + \eta_{ij})$, yields the same estimating equation, but with $d_1 = 1/\theta$ which can differ from 1. Thus, we do not impose $d_1 = 1$ in the regression.

5.2 Editorial Responses to Prolific Referees

The second key deviation from the benchmark of citation maximization is with respect to reliance on the referees: editors appear to give 20% higher weight to the recommendations of more prolific referees, despite the fact that their recommendations are no more informative about future citations than the recommendations of less prolific referees.

One explanation is *incorrect beliefs*: editors may expect that prolific referees are better judges of quality and therefore give more weight to their opinions. Alternatively, it could reflect a version of the *quiet life* hypothesis: editors may know that prolific referees are no more informative, but they find it costly to ignore their recommendations. We provide evidence on the first interpretation with a survey that elicits, among other questions, beliefs about the informativeness of reports.

5.2.1 Forecasts of Editorial Findings

In the Fall of 2015, in advance of a presentation of this paper, we surveyed a group of editors and associate editors at the *REStud*, and a group of faculty and graduate students at the economics department of the University of Zurich.[43] In the spirit of DellaVigna and Pope (forthcoming), we asked for forecasts of several findings of this project via an 11-question Qualtrics survey. We received 12 responses by editors and associate editors (editors for brevity) at the REStud and 13 faculty and 13 graduate students in Zurich. No draft had been available at the time and these were among the first presentations, making it very unlikely that the respondents could have known about the results.

Table 4 presents the responses to the most relevant questions (not in the same order in which they were asked).[44] Overall, the forecasts by editors and faculty respondents are quite accurate, with an average absolute deviation in percentage points between the correct answer and the average forecast of 6.4 (editors) and 5.1 (faculty). In comparison, graduate students have a deviation of 8.8.

The first two questions elicit a measure of how well editors and other economists understand the uncertainty in forecasting citations at the desk-reject stage overall ("*What percent of desk-rejected papers end up in the top 5 percent of citations (by the Google Scholar measure)?*"), and for submissions of prolific authors ("*Consider all submissions with at least one 'prominent' coauthor that are desk- rejected. What percent of these papers end up in the top 5 percent of citations?*").[45] The responses by the editors suggest that they are aware that at the desk-reject stage they set a higher bar for papers by prominent authors. The responses of Zurich faculty, however, suggest that they do not anticipate this higher bar. On one of our key findings there is therefore significant disagreement.

Next, we ask for a forecast of how predictive referee recommendations are of citations: "*How much higher is the percentile citation if a referee recommendation is positive versus if it is negative (for papers with 3 reports)?*" and the ask the same question for reports of "*prominent referees*". Recall that we find no indication that more prolific authors are better able to forecast citations. Nevertheless, REStud editors expect the recommendations of prolific referees to be nearly 40 percent more informative. This average does not reflect an outlier forecast: 9 out of 12 editors expect prolific

[43] The survey sent to the REStud editors refers only to the REStud, while the Zurich survey refers to all 4 journals.
[44] The full survey is in the Online Appendix.
[45] We define prominent as having "*published at least 4 papers in the 5 years before submission in 35 high-impact economic journals*". We did not ask a parallel question for the R&R decision.

referees to be more informative. We find a similar pattern for faculty and graduate students at Zurich. This supports the hypothesis that incorrect beliefs induce additional reliance on prolific referees.

We also elicit a measure of how well editors are able to forecast citations at the R&R stage ("*What percent of papers with a Revise-and-Resubmit in the first round end up in the top 5 percent of citations (by the Google Scholar measure)?*"). Editors overestimate their ability to pick top-cited papers at this R&R stage (average forecast of 32.5% versus the actual 18.1%). The faculty and graduate students in Zurich, by comparison, err in the opposite direction.

We also elicit a measure of how closely the editors follow the referees. To keep things simple, we ask for the share of papers with 3 reports that receive an R&R, as a function of the referee recommendations. The respondents appear to have a relatively good understanding of the degree of reliance on the referees, though they appear to underestimate the influence of the referee opinions as a whole. Editors, for example, give a predicted R&R rate of 21.3% for papers with one positive and two negative referee recommendations, while the true share is only 6.4%.

Finally, in the ReStud editor survey we also asked "*Citation-wise, which group of Revise-and-Resubmit decisions do you think does better in terms of later citations? - Papers where the editor follows the referees - Papers where the editor overrules the referees or the referees are split - The same*". Only 6 out of 12 editors give the correct answer (the first one).

There is, thus, interesting variation in the deviations of the forecasts from the observed editorial patterns. We hope that this combined evidence contributes to a more complete understanding of the editorial process among authors, referees, and editors.

6 Conclusion

Editors' decisions over which papers to publish have a major impact 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 to publish. 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 unbiased measure of quality.

This simple model is consistent with several 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 two important deviations from this benchmark. On the referee side, certain paper characteristics are strongly correlated with future citations, controlling for the referee recommendation. This suggests that referees impose higher standards on certain types of papers, or that they discount the future citations for these papers. In particular, referees appear to substantially discount the future citations that will be received by more prolific authors. Editors exhibit a similar penalty in both their revise-and-resubmit decisions and the desk-reject decisions.

We consider two main interpretations for this first deviation. Citations may be inflated measures

of quality for prolific authors, leading referees and editors to discount their citations. Alternatively, citations may be appropriate measures of quality but referees and editors set a lower quality threshold for less prolific authors, perhaps reflecting a desire to help these 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 explanation that referees and editors are effectively easing entry into the discipline for younger and less established authors. Nonetheless, we acknowledge the limitations of this indirect piece of evidence.

The second key deviation is that the editors put more weight on the recommendations of more prolific referees, even though these referees' recommendations are no more predictive of future citations. We consider two main interpretations: editors may have inaccurate beliefs about the informativeness of prolific referees, or their choices may reveal a desire not to disagree with prolific referees. A survey of editors and faculty supports the first interpretation: both editors and faculty expect prolific referees to be more informative.

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 et al. (2016) provide some evidence from a medical journal of favoritism towards prolific authors, a finding different from ours. Another important set of questions concern the initial selection of referees and the dynamic process by which editors decide whether to reach a decision with the reports received so far, wait for more of the original referee(s) to respond, or recruit new referees. We hope that future research will be able to address these and other questions.

References

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

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.

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.

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 Increase 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.

DellaVigna, Stefano, and Devin Pope. Forthcoming. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy.*

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

Ellison, Glenn. 2011. "Is Peer Review in Decline?" *Economic Inquiry*, 49(3): 635-657.

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., 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): 1-23.

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.