Editorial Incentives &

Early-Stage Determinants of Submission Impact

Ayeh Bandeh-Ahmadi[*]

University of Maryland

Abstract

Despite ongoing interest in the role of journals and citation impact on innovation landscapes, the editorial process itself is poorly understood due to privacy concerns and a lack of data; accordingly, there is also a limited understanding of what data should ideally be collected. Through semantic and econometric analysis of editorial databases from five journals covering 2004-2010, including access to paper submissions, revisions, referee reviews, and editorial and referee decisions, I study the early-stage determinants of editorial and referee decisions as well as the information content of referee review. The results both help understand review processes at these journals and identify critical data points that scientific commons could collect and publicize via mechanisms that respect privacy of both authors and referees.

As part of this analysis, I disambiguate referee language that are significant predictors of referee scores versus predictors of eventual citations, finding evidence for instance that papers that have unique datasets or contributions to ongoing debates and government policy, according to referees, receive more citations. I use these findings to develop an estimator of potential impact for all reviewed submissions, published or unpublished. I also measure submissions' textual overlap with past accepted articles in the same journal. Using these metrics, I develop and estimate a model of editorial decision-making over submissions' early-stage impact potential, accuracy, fit within a journal's niche and other characteristics. The results provide evidence of a number of interesting effects: First, they support other economists' hypotheses that there is a great deal of rich information content in referee reviews. Secondly, they suggest that the appearance of favoritism amongst editors who accept a higher share of papers that cite themselves is actually often a reflection of an ability to select for papers in one's own area that have greater citation impact in the long term. Third, I find evidence of an interaction between accuracy and potential for citation impact that suggests it may be much more risky to produce and submit papers that are expected to receive high numbers of citations at certain journals. Finally, I am able to characterize budget constraints on publication space and referee capital.

These results underscore the value of understanding editorial processes across fields and journals for policy-makers, author-researchers, promotion boards, and grant review boards. They also provide some guidance on what types of data good scientific commons on the editorial process could capture, including full texts of referee reviews for journals willing to adopt this requirement or alternatively check boxes for referees to indicate different types of merit (eg.

[*] Please email any comments to: abandeh@umd.edu.

unique data sources or experimental setups, contributions to ongoing debates or policy, etc.), as well as computational methods for identifying what the relevant types of merit are at each journal. The results also suggest that measures of textual similarity to past accepted work as well as keeping information via anonymous flags on citations of referees or editors' own work in a paper are valuable to understanding the incentives at play as well. These metrics can be used to score referees and editors if desired. Altogether, the metrics described here capture a rich amount of information about editorial processes and are simple to automatically compute given modern software technologies. As a result, they are important to consider for developing not just commons for editorial decisions but for any type of scientific review which is designed to encourage scholarship and transparency.

JEL Codes: A14, C81, D0, D2, D71, D8

Keywords: editorial decisions, referee reviews, information asymmetries, collective decision-making, textual analysis, journals, R&D, innovation, mechanism design, scientific commons

1 Introduction

As gatekeepers who decide what new ideas in their areas of expertise are worthy of publication, editors face a variety of incentives. Journal editors may seek to maximize their journal's ranking on impact factor, which is a measure of the frequency with which the average article published in the journal receives citations over a certain period of time. Higher impact factors often mean that more researchers pay attention to the journal; in turn these may lead to greater prestige and influence for the editors as well as higher profits via institutional subscriptions. Pressure to increase profits from publication could also encourage editors to broaden the readership and niche of the journal or both; journals may be able to use larger numbers of published articles per issue to justify a higher subscription fee. On the other hand, pressure to increase profits could also push editors to spend less money on publication costs in the form of fewer articles and limited budgets to cover time and other expenses. While a broader niche may appeal to a broader readership and bring a broader pool of articles from which to select quality publications, a more specialized niche may lower the editorial workload involved in processing large numbers of submissions and increase attention and quality of submissions from researchers within the specialized area. In addition, editors have an incentive to publish articles that are accurate because they face potential embarassment from colleagues if they publish research that is later refuted.

Therefore, a submission's technical accuracy and merit, potential for impact, and fit within the journal's niche are all potentially relevant to editors' decisions. In addition, since editors are usually either active or recent researchers in their journals' niches, there may be personal spillovers an editor expects to receive from publishing a submission. Yet, while there have been a number of economic studies of outcomes and durations of the editorial review process, little work has gone into understanding the relative importance of and interplay between these factors in determining editorial incentives. In addition, there is little existing work highlighting how editors gather and interpret information on each of these fronts.

The limited literature with access to data on both accepted and rejected submissions, such as Abreveya and Hamermesh (2009) and Hamermesh and Oster (1998), describes secondary characteristics of papers that are more likely to get accepted (authors' gender, institution rank, age, et cetera) and in some cases offers insight into which editorial practices (eg. single- or double-blind) are more biased[1] on these characteristics (Blank, 1991). Cherkashin et al (2009) even make use of data on rejected papers that were eventually published elsewhere to evaluate how good of a job editorial bias does at selecting quality papers, with quality defined as greater citation impact. What is still missing from this discussion is an examination of whether impact is actually the main driver of editorial decisions and what role other factors play.

While various explanations have been hypothesized for all these biases, very few studies have tried to directly analyze the determinants of editorial decisions.

Ellison (2002), Laband (1990), and Laband and Piette (1994) are notable but very limited exceptions. Ellison provides a theory of editorial incentives and referee communication on his way to explaining why editorial lags may have increased over the years. He hypothesizes that editors value quality along two dimensions, i.e. those that are inherent in a paper and those that can be improved with revision. He makes the case that the balance between value placed on the two is a social norm that may change over time. In his model, authors are the only ones who act "non-mechanically;" referees and editors receive perfect signals on each of these two dimensions and select those submissions with the highest combined quality to publish. If referees' expectation regarding the social norm thresholds for publication in their fields comes from personal experience sending their own submissions to journals, and if referees overvalue their own work, then Ellison finds referees may become tougher and demand more revisions over time, lengthening the publication process. He suggests that the most worthwhile extensions of his work would be to examine what happens if referees send/receive noisy signals and if there are editor fixed-effects reflecting more "revision-loving" editors.

Laband (1990) provides survey data characterizing the role of referees in the review process. His findings suggest that referees' letters are very helpful to authors and contain much more information than editors' letters. Accordingly, he makes the case that referees provide much more than a signal to editors about which papers are most promising for publication; rather, they are active contributors to the production of good papers. Laband also provides a striking quote from one of the editors in his study who

[1] I use the term bias broadly here in order to encapsulate both desirable and undesirable selection behavior. Obviously, editors are hired to bring certain bias for selecting high quality papers. They may also have other biases that are not desirable from the perspective of publishers or researchers working within their fields.

laments that he spends too much time trying to improve the quality of “marginal” papers. While this may seem a stunning admission, Laband notes that this editor is only admitting what the data show for every other journal in the study, namely that each publishes a number of papers that go on to receive zero citations. He argues these anecdotes reflect a shortage of high-quality paper submissions at journals, causing editors to rely more heavily on reviewers to help improve marginal submissions to the minimum acceptable threshold for publication. Finally, Laband also suggests that matching submissions to good complimentary referees who can play a productive role in the development of those assigned submissions may be one of the primary roles of an editor. [2]

Laband and Piette (1994) look at citations data on published papers to show that when editors publish work by past coauthors or graduate students, it usually though not always goes on to have higher impact, suggesting that editors use their networks to identify and capture better research and improve efficiency in the market for scientific knowledge.

A main reason for the dearth of research on editorial incentives is that it has traditionally been difficult to measure a submissions’ fit to a journal’s niche, the amount of impact editors expect it to have, the level of new or innovative content it contains, and even its technical merit or accuracy. Yet, we know that journals pay attention to their impact ratings, that they exist in niches with varying degrees of readership and communication overlap, and that journal editors are embarrassed by having published subsequently-refuted findings. So although journals may not directly store metrics on each of these factors, we can surmise that since editors use the review process to learn about these factors, information on each these factors must be captured in editorial databases during this process. In fact, as Laband (1990) suggests, referee reports may be a good place to look.

In this paper, I set out to decompose editorial incentives on notional submission quality beyond what has been presented in previous literature. I seek to address how journals decide which submissions to publish, including the role of accuracy, potential impact, and fit amongst editorial incentives. I consider what kinds of signals referees send editors and how editors interpret this information. I also look for evidence of binding budget constraints on publication space and referee capital as well as evidence

[2] This observation was also made by editors participating in this study.

of personal bias towards submissions that cite research by journal coeditors. In order to address these questions, I develop new text-based metrics from manuscripts and referee reports which I combine with data on authors' countries, submissions' JEL codes and coeditor assignments.

My findings provide empirical validation and significant extensions of a number of hypotheses presented by Ellison (2002), Laband (1990) and Laband and Piette (1994). In particular, I demonstrate that initial referee reports contain a great deal more information about potential impact than is captured by the scores referees assign. A consequence of this finding is that it is possible to disambiguate referee insights on potential impact from other referee characterizations of quality, thereby providing a rich, new source of data into the determinants of impact. I build a basic version of such a model to instrument for potential impact of published and unpublished submissions while addressing most if not all endogeneity concerns.

I find that this instrument for potential impact enters an editor's decision principally through interaction with the number of negative reviews a submission receives, suggesting first that editors are especially averse to the risk from publishing refutable results that could garner more press. Secondly, this implies that editors are not constrained with too many high-quality submissions, just as Laband's editorial anecdotes suggest.

Furthermore, consistent with Ellison's hypothesis, I show that referees at many but not all journals do in fact become more harsh over both time and with age. Finally, I am able to characterize editor-specific fixed effects that could reflect better matching ability to referees and/or revision-loving behavior that Ellison's theory describes.

Besides the ongoing discussion of the editorial review process, the analyses presented here draw on the economic literature on information assymetry. Akerlof (1970) most famously characterizes signalling and screening as two tools available for distinguishing the quality of goods in the presence of incomplete information. Both of these are available to editors. The questions here are what information is present in referees' signals to editors, how do editor's interpret this information, and what types of screening are editors able to induce either through the revision process and/or their personal networks.

In addition, these analyses contribute to the literature on understanding the determinants of innovation, particularly at the individual level. Goyal et al study characteristics of individuals (such as early mentorship by other high-productivity individuals) that make them more likely to become superstars (defined as having many publications). This investigation compliments existing work to the extent that it contributes to an understanding of the determinants of editorial decision and of eventual citation impact of published work.

Finally, to the extent that papers and authors can be considered goods and firms that are in competition with one another, this study sheds light on the nature of spillovers between them and the role of niches and specialization not just in the organization of journals but in the development and spread of innovation. Journals face interesting multi-sided markets in that demand for publications comes from both readers and authors. The complicating factor is that unlike traditional multi-sided markets, the readers, authors and even referees are subject to a great degree of overlap. The potential exists not just for collusion in the form of favoritism but also for complimentarities and specialization in the production of knowledge. The findings here about personal spillovers between editors and authors and about spillovers between articles with higher fit help to characterize the organization of these players.

The remainder of this paper is organized into a number of sections. Section 2 gives an overview of the editorial data that provides the basis for this paper. In Section 3, I discuss textual analysis tools to examine a journal's subject niche as well as methods to measure submission fit to said niche. In section 4, I develop and evaluate results from a basic model of citation impact based on referee language and other factors that can be applied to both published and unpublished submissions. In Section 5, I study the determinants of referee decisions at the same journal, evaluating the role of various factors including referee taste for impact to develop a clearer understanding of what referee signals communicate. Section 6 outlines a general model of editorial decision-making informed by the submission data, referee input and textual metrics described in previous sections. Section 7 gives an in-depth econometric assessment of editorial decisions, incorporating analysis of referee reports and citation impact together with a more basic assessment of editorial decisions across all five journals. Section 8 provides a discussion of the results and section 9 concludes.

2 Editorial Databases

The data used in this study come from several scientific journals that have granted permission and secure access to their editorial databases as well as from public sources of citation data online, including REPEC. These databases include complete information used in the manuscript submission process including submissions, decisions, and all correspondence between authors and editors.

In order to respect the privacy of the journals, their identities and those of their editors, authors and referees are treated with strict confidentiality. Certain additional data, such as JEL codes, are also suppressed in the analysis below in order to respect this confidentiality. Finally, the scope of this work was reviewed and approved by the Institutional Review Board of the University of Maryland to ensure it does not fall under the use of human subject data.

2.1 The Editorial Process

Briefly described, journal submissions are first received by the editor-in-chief at each journal, who then assigns each submission to a co-editor, usually based on co-editor expertise and workload. After reviewing the submission, the co-editor will either make a summary desk rejection or select referees and distribute the submissions to them. Each of these referees will either decline or agree to review the submission; referees agreeing to review will provide an evaluation score together with a written referee report outlining some combination of the submission's contributions, shortcomings, suggested changes and a suggested course of action. For submissions that did not receive an initial summary desk reject, the co-editor waits for all referee evaluations to arrive before reviewing them and making a decision. Editors generally attempt to obtain reports and scores from all referees in order to avoid bias towards certain reviewers, but in rare cases exceptions might be made. The editor's decision could be final or may provide an opportunity for the manuscript to go through another round of review and decision following revisions.

For their part, editors choose one of the following decisions for each submitted revision:

Editor Decisions
Summary Reject, No Referee Input
Summary Reject, Referee Input
Reject
Returned for Revision

Conditionally Accepted (Minor Revisions)
Accept

Likewise, referees choose from amongst the following evaluation scores to assign to each manuscript revision they review:

Referee Scores
Definite Reject
Reject
Weak Revise & Resubmit
Revise & Resubmit
Strong Revise & Resubmit
Accept with Revisions
Accept

2.2 Available Data

Altogether the five journals contain 18,241 submissions covering 2004-2010. Of these I have extracted full texts of 6,759 manuscripts. Each editorial database includes JEL codes (where applicable), author names, data on the corresponding authors' addresses and institutions, and an assigned editor ID for each paper submission. It also includes manuscript texts, referee reports, referee evaluations, editorial decisions, and corresponding document receipt dates for each revision of each manuscript.

At the journal for which more extensive analysis described was carried out, 3,391 submissions were received. Of these, I have extracted texts of 2,633 manuscripts and earliest versions of manuscripts for 2,173. In addition, I processed full texts of 2,930 referee reports corresponding to 1,473 manuscripts. Of those referee reports, 458 correspond to initial reviews of 220 submissions that were eventually published. I have been able to identify citation counts for 187 of these publications.

Citations data was collected by hand through REPEC over the course of a week in order to ensure that there was no significant date bias in the number of citations each published paper received.

3 Textual Analysis & Measuring Fit

3.1 Text-based methods

In order to analyze the textual content of manuscript submissions and referee reports, I develop a text parsing algorithm to identify the number of occurrences of every word or phrase, storing the output in the form of term frequency vectors for each document. Because of the different analyses I carry out on each type of document and because manuscripts are much longer than referee reports and therefore more computationally intensive to process, I use slightly different methods for processing each of these corpuses of documents.

For manuscripts, I store frequencies of three-word terms of at least 12 characters occuring in at least two documents. These term frequency vectors are used to calculate textual similarities between submissions and accepted submissions from the previous year, as described in section 4.1.1 below. In addition, I store all occurrences of the last names of the journals' primary co-editors in both earliest and latest revisions of all manuscripts for further econometric analysis of decisions.

3.1.1 Measuring Textual Similarity

Measures of textual similarity have been widely used by computer scientists for a variety of purposes, from search-engine algorithms to image recognition. Hoberg and Phillips (2010) use it on the texts of thousands of 10-K filings to show that firms with higher similarities are not only more likely to merge but also show higher post-merger increases in revenue. Here, I use similarity to measure how closely related each submission is to past accepted work at the journal, which I will employ as a proxy for fit.

I measure textual similarity between documents as the cosine similarity between their term-frequency vectors (Salton and McGill, 1983). Cosine similarity is one of the most commonly applied measures used in computational linguistics and is defined as the

$$similarity_{kl} = cos\left(\theta_{kl}\right) = \frac{\vec{t}_k \cdot \vec{t}_l}{\left\|\vec{t}_k\right\|\left\|\vec{t}_l\right\|} = \frac{\sum_i t_{ki} t_{li}}{\sqrt{\sum_i \left(t_{ki}^{\ 2}\right)}\sqrt{\sum_i \left(t_{li}^{\ 2}\right)}}$$

length-normalized dot product of a pair of vectors. It captures the amount of textual overlap between the text in two documents with term frequency vectors t_k and t_l.

Note that because of the length-normalization in the formula above, similarity between each pair of papers is bounded between zero (no textual overlap) and 1 (all terms in each document occur in the other).

Appendix A shows the distribution of cosine similarity across all pairs of paper submissions and accepted paper submissions received in 2009 at Journals 1 through 5. These similarities are calculated based on all one-word terms that occur in at least two manuscripts. As these graphs show, the average cosine similarity between two submissions to the same journal is between 4.9%-8.6% depending on the journal, with the average being slightly higher when limited only to accepted papers.

3.1.2 Constructing a Measure of Fit

As a result of performing textual analyses of accepted manuscripts detailed above, I am able to examine what constitutes a journal's niche within this framework. The first step is to investigate the relationship between submissions and accepted papers. Using the cosine similarity metric, it is possible to visualize the distribution of relationships between paper submissions in a network layout, such as the one in Appendix B. Here papers submissions received at The journal in 2009 are connected if they have a cosine similarity greater than 0.18[3]. The spring-embedded layout algorithm used (Kamada Kawai, 1988) aims to place papers that have more relations in common closer together. While there are many different kinds of layout algorithms for visualizing networks, the graph in Appendix B helps to visualize and find where clusters exist as well as what kinds of cross-cutting relationships occur amongst different groups of papers. It is much more easy to see that accepted papers occur across a broad range

[3] Since cosine similarity scores are relative to the corpus and types of tems that are counted, this threshold of 0.18 was chosen based on a reading of documents to determine the threshold above which papers seemed quite related to one another; it also corresponds to the cosine similarity between Ronald Coase's "The Nature of the Firm" and Oliver Hart and John Moore's 1990 "Property Rights & the Nature of the Firm."

of these clusters than it would be from a pair-wise similarity matrix. As this graph shows, there is a rich set of relationships amongst journal submissions and accepted submissions that is important to take into account when defining the journal's niche.

In order to accommodate the fact that a journal's niche reflects not just a singular paper but rather a corpus of somewhat loosely related past publications, I construct a proxy for a submission's fit within a journal's niche by averaging the submission's similarity across all papers accepted at the journal from the previous year. Because pair-wise similarities are restricted between 0 and 1, this method provides a measure of relation to the niche recently defined and covered by the journal that also falls into this range, as discussed in Section 4.

$$py\_accepted\_sim_k = \sum_l similarity_{kl}$$
$$l = set\ of\ accepted\ \ papers\ \ from\ previous\ year$$

Appendix C shows the distribution of average similarity to previous year accepted papers across all submissions received in 2009 at one journal. As this graph shows, more papers are only slightly related to the previous year's accepted papers while a few are much more broadly related. These similarities are calculated based on all three-word terms of at least 12 characters that occur in at least two papers. These restrictions were chosen because of computational limits; in order to be consistent with the measure presented in Appendix A, I expect to update this metric.

4 Modeling Impact using Referee Language & Other Determinants

For the next step in building this framework, I develop a method for modeling the potential impact of individual journal submissions by investigating the correlation between various determinants and citation outcomes. First, I test the hypothesis that language used in referee reports contain human-coded characteristics not otherwise observable by identifying specific language that are correlated to referee-assigned scores. Second, I incorporate only the most significant referee language to build a model predicting impact of any individual submission. Last, I examine the effect that other determinants such as editor preferences and bias may have on submission impact.

4.1 Identifying Referee Language Correlated with Score and/or Impact

In order to identify specific referee language that is a significant predictor of citation outcomes at The journal, I regress citations on term frequencies from the earliest round of referee reports available for published submissions. Casting a broad net on language that could be meaningful, I examine one- to three-word phrases of at least four characters that occur in these referee reports. By restricting this process to only terms applied in at least 15 referee reports and by using only one term at time as an explantory variable, I am able to avoid any effects that come from singling out individual papers. I also control for date in each regression, since later publications enjoy less time in which to build citations.[4]

In addition to regressing citations one-by-one on referee terms, I also conduct a similar set of regressions of referee-assigned scores on referee terms in order to evaluate to what extent referee scores capture anticipated citations.

In each set of regressions, the significance of each referee term can be interpreted based on its p-value and coefficient. The two righthand plots in Appendix D shows these terms plotted based on their score coefficients and citation impact coefficients (the upper plots are graphed based on rank of coefficients whereas the lower plots show actual coefficient values). The lefthand plots shows these terms plotted based on their significance (i.e., p-values of their coefficients on score and citation impact). Again, the upper plot shows rank (p-value) along both axes while the lower plot shows actual coefficients. Plotting the rank of coefficients and p-values is similar to plotting these values in a log-log format in that it makes it easier to display very large and very small values on one plot.

The p-values of these terms in both regressions run the gamut from less than 10^{-10} to nearly 1. Likewise, the coefficients vary from less than -150 to 100 on score and from -5 to almost 30 on citations. Given the ranges covered by these coefficients' p-values, and the fact that all terms were applied in at least 15 referee reports, there are clearly some terms that are highly significant predictors of impact and terms that are highly significant predictors of scores.

[4] Interestingly, I find that when modeling impact, estimating a coefficient on the date on which the submission was received is more significant than the date on which the submission was accepted. This may reflect a trend in certain fields such as economics and finance, of authors broadcasting working versions of papers which can get cited, sometimes long before they get accepted at any journal.

In fact, it is quite telling that there is virtually no correlation ($R^2=0.025$) between how significant a term is in predicting referee score and how significant it is in predicting eventual citations. As a result, we can disambiguate terms that are significant predictors of the referee's score from those that are significant predictors of eventual impact to build an instrumental variable for eventual impact without endogenizing referee score.

As a cautionary note to the reader, textual analyses often results in informative text embedded within noisy text. Furthermore, interpretation of these terms are complicated by the fact that informative text may seem to be meaningless at first glance. For instance, the highly significant term "from to" corresponds to language constructs such as occurrences of "from [year] to [year]," which is frequently used to characterize unique datasets. As with other kinds of natural language processing, the key is to identify enough datapoints in order to be able to extract a meaningful signal from the noise. In this paper, I chose a simple model to do so as described below.

4.2 A Basic Model of Predicted Impact Based on Referee Language

In this section, I develop a basic model of impact based on referee language taking the most significant referee terms for predicting impact (i.e., those with the highest p-values from the one-term-at-a-time regressions above) and regress citations on all of them together, as well as referee score and date of submission receipt. Doing so yields a fit with an adjusted R^2 of 0.63 compared to an adjusted R^2 of 0.05 when only referee score and date are used as predictors of citations. Although it is likely that refining this model of language predictors may yield significantly better fits, I proceed with building the framework using this fit to the data.

Because many textual terms' frequency and occurence are correlated with each other, two terms may capture much of the same information. Putting these terms in a regression together results in the coefficients on some terms no longer being significant. To reduce dimensionality, those terms that are no longer significant can be removed from the regression, resulting in an adjusted R^2 of 0.6. The results are shown in Table E. Again, each of the terms included here is highly significant on its own at the 1% level or better and has been applied in at least 15 referee reports.

4.3 Referee Language Predictors of Impact

The most significant terms for predicting citations, scores, and both citations and scores are listed in Appendix F. These suggest there are a variety of different determininants correlated with larger numbers of citations. Some determinants of citations are topical, reflected in terms such as "wage," and "interest rate." Others characterize the nature of conversation a submission contributes to; "debate," and "government" are examples of such referee language. Still others characterize a pure opinion, such as "opinion," or "does not." Finally, some such as "from [year] to [year]" characterize an aspect of a paper that is valuable, such as a unique dataset. In future work, it will be valuable to spend more time examining each of these determinants, how referee language is applied across referee reports, which terms covary with each other, and to study how corelated these factors are across journals.

This framework combined with this rich dataset potentially provides a way to examine specifically how determinants contribute to submission impact. Furthermore, it may provide a way to better understand how these determinants relate to identifiable characteristics in submissions, potentially providing a way to characterize innovation within each submission.

4.4 Other Determinants of Publication Impact

The results in Appendix G and H characterize other determinants of publication impact, including paper characteristics and editor effects respectively.

Length: Beyond referee language predictors of impact, the positive coefficients on both earliest and latest version manuscript length (measured in words) in Appendix I seem to suggest that longer papers go on to garner more citations -- three more citations per one thousand words in earliest submitted versions of manuscripts and an additional 0.5 more citations per thousand words in the last revision of the manuscript. One might speculate that longer earlier versions reflect greater thoughtfulness and effort on the part of authors or simply more content that may we worthy of eventual citations.

Editor Dummies: Editor dummies characterized in Appendix J vary in magnitude and significance. This suggests either some editors do a better job of selecting more impactful papers, place higher weight on impact over fit/accuracy, or that have less buzz-worthy niches. Cherkashin et al (2009) give a further explanation for this latter case, suggesting that the editor-in-chief may cherry pick better papers for his/herself or favored editors. In the case of This journal, however, the coefficient on the editor-in-chief's dummy variable is near the middle of the range for all editors, suggesting that he/she is at least not cherry picking better articles for his/herself.

Mentions of Co-editors' Names: Mentions of co-editors' names in final versions of manuscripts seem to have different effects on eventual citations. The sign and significance of these effects vary across editors as shown in Appendix J. The effect is only significant for those two editors responsible for the largest number of manuscripts. For the editor-in-chief, the effect is positive and highly significant at the 1% level, suggesting papers closer to his/her niche get more visibility, perhaps because a large number of readers interested in that line of research pay attention to the journal. On the other hand, the effect is negative and still significant at the 5% level for Editor 1, who is responsible for editing the second largest number of manuscripts.

Interactions between Co-editors and Mentions of Co-editor Names: These are not significant, except in the case of the editor-in-chief. This suggests that he/she has a lower bar for accepting papers mentioning his/her own name. If this is true, then the editor-in-chief's editing is not likely responsible for the higher number of citations received by papers assigned to and mentioning him/her. Instead, this seems to suggest that there is a readership effect and that this effect does not extend beyond the editor-in-chief to his/her co-editors -- in other words, the choice of editor-in-chief is an important signal for establishing the journal's readership. An alternative but less plausible explanation also exists; it is possible that the publishers have a high incentive for choosing an editor-in-chief that is in line with their existing readership; nonetheless, the editor-in-chief holds most of the responsibility for making decisions regarding what to publish and therefore still has to be responsible for maintaining that readership. Therefore, in either case, it seems this person serves as a figurehead in attracting and maintaining journal readers.

5 Investigating Trends in Referee Behavior

This section reports results from analyses of referee decisions to agree to review papers and from subsequent referee scoring behavior.

Appendix I.1 characterizes a binomial logit on referee decision to review a manuscript. The results show that referees at Journals 1 through 4 are less likely to accept a referee request with more past reviews they have submitted to the journal.

They suggest that whatever the returns to refereeing are, they have diminishing marginal returns. Furthermore, they also suggest editors at Journals 1 through 4 are somewhat constrained in the amount of referee capital they have.

Results from a multinomial logit on referee-assigned scores across Journals 1 through 5 are shown in Appendix I.2. The data suggests that referees at Journals 1 through 3 and Journal 5 generally assign lower scores over the course of time when controlling for number of past revisions. The exception is Journal 4 where referees grade slightly but significantly more generously with time. At all journals except Journal 5, the negative coefficient on total reviews suggests that referees grade more harshly later in their tenures as well, controlling for date.

Later revisions seem more likely to get accepted at Journals 1, 2, 3 and 5, signaling that referees, editors and authors are doing a good job of selecting the right submissions to revise-and-resubmit. In contrast, results at Journal 4 show referees grading slightly more generously with time, and later revisions are significantly less likely to get accepted, indicating many more revise-and-resubmits awarded than eventually get published.

Coefficients on length are significant and negative in all journals. This could mean that referees prefer shorter papers, or it could reflect greater effort devoted to editing down final versions of papers that go on to be accepted. Subsequent results on editorial preference for longer papers help to detangle these alternatives, providing support for the former explanation.

Finally, coefficients on average similarity to previous years' accepted papers vary in sign and significance across journals, suggesting that at some journals referees may take a greater interest in whether a submission fits the relevant niche, while at others referees may be concerned with broadening the journal's niche. These results suggest that it is important to interpret the results regarding referee preferences in a journal-specific context.

Appendix J gives a more in-depth look at referee evaluations at One journal. As discussed above, the significant and negative coefficient on submission receipt date reflects greater selectivity by referees over time now that there is no control for revision number or total reviews.

Coefficients on editor dummy variables are all positive and significant but vary in magnitude, suggesting quality of papers may vary across editors and/or their field specialties. An alternative explanation is that put forth by Laband that some editors consistently do a better job of matching referees who are able to improve the quality of marginal papers. It is also possible that referees in certain fields are generally more lenient than others.

There is a very small but significant effect of the referee language instrumental variable for potential impact (E(citations)) that appears to dissipate when not controlling for revision. This suggest that expected impact plays a minor role in referee score. This finding is consistent with findings in the previous section that there is low correlation between language that predicts referee score and langauge that predicts eventual citations.

6 A General Model of Editorial Decision-Making

Having characterized all the different sources of data on signals available to the editor, I will present a model for characterizing editorial incentives over all of these factors. I consider a setting in which the editor's problem is to pick the set of papers that maximize his expected utility subject to constraints on time and the number of papers that can be published. If we assume that an editor's expected utility is well-aligned with that of the journal, given the discussion above, there are a number of factors that could enter into his utility function: these could include each submission's potential impact, its fit to the journal's niche, and its accuracy or possibility of diminished reputation from exposed inaccuracies. In addition to factors that directly affect the journal, an editor may have personal bias based on a submission's relationship to his own work and/or the opportunity a submission presents for bringing more citations to himself.

First, I define a journal payoff function for each submission which an editor publishes: $\Pi_\mathrm{p}(a, i, f)$, where a captures the published article's technical accuracy and merit, i reflects its potential for generating interest and impact, and f characterizes its fit to the journal's niche.

Under this model there appear to be ready bounds on the functional form of Π and on each of the other variables. For example, since it is difficult to imagine an article receiving a negative amount of interest in much the same way that it is difficult to imagine a negative number of citations, I define the range for i from zero to infinite.

Although an article may very well receive “negative” press, that negativity would reflect poor accuracy or deficiencies in technical merit rather than of the magnitude of press.

Similarly, I define the range of fit bounded from zero to 1. This metric is consistent with the average cosine similarity metric presented in Section 3 and also with the general notion that negative or infinite topical fit would be nonsensical; either a submission has some relevance to the content the journal publishes or it is not relevant at all.

Lastly, for the purposes of this model, I assert that the collection of statements within an article to either be true or untrue, so that accuracy too is bounded from −1 reflecting wholly unsubstantiated or untrue remarks to 1 reflecting fully substantiated and true remarks. To further simplify the model, I consider the range of a further to a discrete space like $\{-1,1\}$ with $a = 1$ reflects sufficiently substantiated work and $a = -1$ reflects insufficiently substantiated work. This latter specification could characterize a framework where the a high share of negative referee scores assigned to a submission reflect a signal of poor accuracy whereas a high share of positive scores may signal high accuracy. This is hypothesis we will assess given the model results.

Thus, for certain specifications of the payoff function, such as a Cobb-Douglas form: $\Pi_{\mathrm{p}} = a^{\alpha} i^{\beta} f^{\gamma}$, a submission would result in negative or positive payoffs reflecting the sign of the technical accuracy while magnitude reflects paper fit and impact. Given such a functional form, papers with greater potential for generating interest (i.e., higher impact papers) would carry more potential risk of damage. Further, we would expect an interaction effect between accuracy and paper impact. Higher impact papers might carry longer review times and/or more rounds of revision unless these submissions also tend to be more accurate.

In addition to the payoff that comes with each article a journal publishes, there are also costs associated with publication in the form of editorial time, reviewer time, and publication financing. All editors who gave permission to use their data in this study remarked on the increasing burdens on their time and that of their reviewers. These anecdotes together with other work characterizing increasing review durations such as Ellison (2002) suggests that editors face a limited supply of referee capital, in this case defined as colleagues’ time on which they can draw during their tenure to ask for help refereeing. Authors are often able to suggest referees for their work, but since the number of referees with the appropriate expertise is limited, the suggested reviewers too often overlap with past referees at the journal.

Finally, editors face physical and budgetary limits on the numbers and length of papers they can publish in each issue in addition to the limits on time available to review, revise, and publish these papers. As a result, editors must act strategically in allocating their time and reviewer capital; they choose which papers to desk reject, how many reviewers to assign, and after how many rounds to accept or reject a paper accordingly. Editorial bias towards shorter papers could reflect a binding constraint on space for publication at a journal.

Rather than begin by estimating a structural model in which the form and interactions between all of these editorial incentives, timing and decisions are known, I begin with a very general framework for evaluating editorial incentives and attempt to characterize as much as possible about them. These insights can then inform and make it possible to develop a good structural model of editor decision.

For the purposes of this model, I assume that editors internalize the functional form of payoff and estimate the expected publication payoff for each submission based on the information they gather during the review process. This information set consists of noisy signals editors receive on the accuracy and potential impact based on their own reading of a submission and the input of referees. I also assume that fit is *ex ante* observable by editors and referees just as it is for us, so editors do not face a signal extraction problem over fit:

$$E(\Pi_{\mathrm{p}}|s_a, s_i) = \iint_{a,i} \Pi_{\mathrm{p}}(a,i,f)\, A(s_a,a)\, I(s_i,i)\, da\, di$$

Here $A(s_a, a)$ is the editor's prior distribution of accuracy given collective accuracy signal s_a and $I(s_i, i)$ the editor's prior distribution of potential impact given collective impact signal s_i immediately before making a final editorial decision. Making the assumption that any interaction effects between concurrent submissions can be captured in the signals received by the editor and in fit with recently accepted articles allows us to treat each editorial decision as a separate problem. The editor selects a threshold based on the number of papers that he/she has resources to publish and makes a decision D as follows:

$$D = \begin{cases} accept\ if\ E(\Pi_{\mathrm{p}}) \geq threshold \\ reject\ if\ E(\Pi_{\mathrm{p}}) \geq threshold \end{cases}$$

In practice, editors work with a backlog of accepted submissions that have yet to be published. Because submissions have arrival probabilities which may vary with the academic year, economy or other factors, the size of the backlog could vary as well. Cherkashin *et al* (2009) study this effect finding that larger backlogs correspond to lower acceptance probabilities resulting in higher thresholds. This follows directly given that larger backlogs give editors the opportunity to be more selective. As it stands, the version of the model presented here does not take backlog into account directly, although it does allow for general date effects that would capture persistent trends in backlog and selectivity. In future versions of this model, I plan on incorporating an explicit duration model that accounts for the backlog in manuscripts.

I use empirical methods described in the preceding sections to further inform the parameters in and validity of this framework. I use the predicted number of potential citations from the model in Section 4 as an instrumental variable for potential impact of published and unpublished papers, i.e. as a proxy for s_i. Given results from Section 4 and 5, suggesting that only a tiny fraction of referee scores are explained by eventual impact and by a submission's fit, I proceed under the hypothesis that referee score must primarily capture accuracy or technical merit of a paper and use it as a proxy for s_a.[5] I use the metrics described in Section 3 as a proxy for fit, f. Under the assumption that referee scores reflect accuracy, I evaluate the interaction between these signals in editors' decisions. And furthermore, I also consider potential additional effects of self-citations for editors.

7 Results from Editorial Decision Analysis

Appendix K-N characterize the relationship between referee decisions to agree to review manuscripts and referee evaluations across all five journals.

Revision: Editors are more likely to accept papers in later revisions. This is presumably because those selected for further revisions are the most promising and despite the fact that they seem to receive fewer citations.

Potential Impact: has a small positive, significant effect. However, this effect disappears when an interaction between potential impact and the number of negative

[5] This is consistent with the assumption that editors do not solicit feedback on fit since it is *ex ante* observable and with my findings that referee scores only capture 5% of variation amongst citations of eventually published submissions.

reviews in the earliest round is added in, suggesting that potential impact primarily enters the editor's decision via the editor's evaluation of risk exposure from publishing a potentially inaccurate paper. The interaction between the number of positive reviews in the earliest round and potential impact is not significant, sugesting that either the editor is not accounting for impact (unlikely) or the editor is not constrained by too little space for publishing all high impact papers -- i.e the editors can not afford to be selective on impact but nonetheless are selective against potentially damaging papers. This finding supports using a functional form similar to the proposed Cobb Douglas function in which there is interaction between impact and accuracy in the editor's decision. These results could potentially vary widely across journals, particularly in journals such as *Science* where the potential for high impact work being submitted is quite high.

Referee Evaluation: Likewise, the number of negative evaluations figures much more prominently into editor decision (that is, carries a more negative coefficient) than the number of positive evaluations, although both are significant. This effect becomes more pronounced in later rounds, presumably because more information or stronger signals about the manuscript are revealed in each round.

Fit: The data here suggests that fit is not a significant factor when controlling for other paper characteristics when considering only three-word terms of at least 12 characters, but it is quite significant at Journals 1, 2, and 4 (and insignificant at Journals 3 and 5) when considering all one-word terms occuring across all journals.

Length: Editors show a positive preference for length at all journals that appears to disappear when expected potential impact is accounted for at This journal. This suggests that longer papers are indeed more well-thought out or contain more results worthy of citations that editors pick up on via the referee signal.

Editor Effect: Dummy variables for editors are negative for all referees except one but vary instead in sign and magnitude. Some are highly significant; others, insignificant. This variation suggests that either some receive better papers because of their specialties or that they may have different quotas or standards.

Editor Citation Interactions: Of the interactions between editor dummy variables and mentions of the assigned editor's name in the final version of the manuscript, only the interaction for one editor is significant. This means not only Editor 1 is more likely to accept papers with his/her name in the last version but also those papers also go on to get cited more, though the latter effect is not highly significant. This suggests that this editor may be attracting more better work to the journal within his/her personal niche, making the case that the choice of co-editors is also significant to the extent of their niches.

8 Discussion

The results shed light on several different aspects of the editorial and publication process. As discussed, the findings provide empirical validation and significant extensions of a number of hypotheses presented by Ellison (2002), Laband (1990) and Laband and Piette (1994).

The findings here suggest that referees scores transmit almost no information about paper fit and only a small amount of information about impact. This leaves the assumption that the remaining 95% of referee score describes a paper's accuracy or technical merits. The findings also suggest that the text of referee reviews convey more than referee scores about the potential of a submission to go on to receive higher citations. In other words, referees have a good understanding of what will go on to get citations even though they do not signal this information strongly via scores. Therefore, these results suggest that it may be possible to get at the accuracy of a submission through referee score, by controlling for language that is indicative of higher citations.

I also identify characteristics of papers, at an earlier stage than previously possible, that go on to be cited more. Generally speaking, this is a question that can be addressed to some extent without access to confidential editorial data. For example, Goyal (2011) studied characteristics of authors and their networks that characterize larger numbers of publications and impact-weighted publications. The model I estimate here complements this existing work by estimating impact considering not just submission characteristics but also referee language that are indicative of high or low citations.

Using this model for potential submission impact could provide a powerful new method to investigate questions within editorial decision making. It provides an estimator of impact for all submissions, whether published or not, that can be used as an instrumental variable in subsequent models. Thus, using this instrumental variable makes it possible to study what information referee scores transmit and how much of a role impact plays in editorial decisions. Also, to the extent that we can use referee scores to control for the higher quality of papers that are eventually published, this instrumental variable could allow us in subsequent work to measure each journal's role in obtaining greater exposure for each article it publishes, disambiguating that effect from each article's inherent technical merit. In addition, the referee language that is identified as significant in predicting eventual citations is meaningful in itself because these terms capture human-coded characteristics of papers that are not readily measurable using other available metrics or even textual analysis of the papers themselves. Evaluating these characteristics can help us develop a better understanding of the role that journals play in broadcasting research across innovation landscapes. Already, I am able to draw conclusions about the role of access to unique datasets, applications to government work or ongoing debates in garnering more citations for a submitted research paper. Further examination of this unique dataset will surely identify more such characteristics as well as characteristics of work that goes on to get few or no citations. Applying these methods to journals in other fields and comparing the results will yield insights about incentives across different fields as well.

8.1 Determinants of Editorial Decision

I find that this estimate for potential impact enters editors' decision principally through an interaction effect with the number of negative scores a submission receives, suggesting that editors are indeed worried about publishing high-impact work that may receive lower citations. Furthermore, this result suggests that editors are not constrained with too many high-quality submissions just as Laband's editorial anecdotes suggest.

What this means for authors submitting papers is that if results are controversial, unique, applicable to ongoing debates or otherwise likely to yield more attention, then it is necessary to put extra effort into ensuring accurate results and clearly demonstrating that accuracy. Alternatively, an author submitting marginal results might be able to successfully progress through the review process by making the work less attention-grabbing in order to fill editors' need for more papers without attracting the kind of attention that editors are risk-averse to. In future work it is worth examining whether these potentially high-impact papers (both those that actually go on to receive more citations and those rejected ones for which referee language predicts more citations) actually receive more editorial attention measured in review times, number of assigned referees, and number of referee rounds.

Additionally, editors seem increasingly sensitive to referee scores throughout the editorial process, suggesting a great deal of weight is placed on accuracy.

Finally, the findings suggest that some countries and JEL codes are more likely to receive acceptances.

8.2 Determinants of Impact

The results here suggest that the answer to the question of why do some papers receive more citations than others is quite complex. It seems that papers may receive more citations due to their topical relationship to the journal and/or its editor but that relevance to hotly debated topics, government applications, unique features such as datasets, as well as other characteristics such as a reviewer simply finding the work "very important" can all correspond to higher eventual impact.

As with other kinds of natural language processing, the key is to identify enough datapoints in order to be able to extract a meaningful signal from the noise. In this paper, I chose a simple model for my initial analysis. Nonetheless, it is worth spending some time more carefully examining the language that appears and other metrics besides those used here for estimating the potential impact of a paper. In particular, it would help to measure citations on a logarithmic or rank scale to account for the fact that they known to have a power-law distribution. There may also be better ways for incorporating a larger number of terms found to be meaningful by increasing the number of referee reports terms must appear in to be counted towards this model or by simply counting the number of positive and negative words that appear.

8.3 Spillovers

There is evidence that published papers closer to a journal's niche get more impact citations, although this and other conclusions based on citation impact are worth evaluating and comparing across other journals. Papers mentioning an editor's name in the first version are more likely to eventually get accepted, when correcting for impact, which is consistent with the finding that they go on to get fewer citations. Editors are harder on (i.e., more likely to reject) papers that cite them. Nonetheless, interaction terms provide evidence that some editors differentiate more (i.e., select higher impact papers) from those that mention their names. Also, some editors appear to choose harder referees or give more desk rejects than others. All of this suggests that co-editors have different styles in the amount of work they are willing to take on and in whether they are harder on papers that mention them. It may be worthwhile for journals to keep such metrics on co-editor's types.

Furthermore, consistent with Ellison's hypothesis, I show that referees at many but not all journals do in fact become more harsh over both time and with age. Finally, I am able to characterize editor-specific fixed effects that could reflect better matching ability to referees and/or revision-loving behavior that Ellison's theory describes.

9 Conclusion

The findings here characterize some new insights about editorial incentives, spillovers between the players involved in the editorial process, and about the determinants of citation impact. By using a broader set of editorial data together with new textual metrics, I am able to show that editors' greatest incentives are to avoid publishing highly-visible potentially refutable work. At least at some journals, there is no binding constraint from too many high-impact papers being submitted, so authors who are able to demonstrate marginal contributions can get published although their work typically does receive lower citations.

Secondly, this work opens up a new source for identifying previously immeasurable characteristics of research which are tied to higher citations and scores. Broadly speaking, the methods applied to develop metrics on referee language here could also be applied to any other source that contains textual and numeric data, for example employee hiring or salary reviews, insurance risk assessments, real estate appraisals, financial filings or legal reviews.

In future work, it will be valuable to spend more time examining each of these determinants, how referee language is applied across referee reports, which terms covary with each other, and to study how corelated these factors are across journals. This rich dataset potentially provides a way to examine specifically how determinants contribute to submission impact. Furthermore, it may provide a way to better understand how these determinants relate to identifiable characteristics in submissions, potentially providing a way to characterize innovation within each submission.

Finally, this work identifies both positive and negative spillovers that can occur when submissions are closely related to and mention an editor's own work. Journals can use the metrics provided herein to monitor the nature of these spillovers and give feedback to editors accordingly.

10 References

Abrevaya , Jason and Hamermesh, Daniel S., “Charity and Favoritism in the Field: are Female Economists Nicer (to Each Other)?” (May 2010). NBER Working Paper Series, Vol. w15972, 2010. Available at SSRN: http://ssrn.com/abstract=1601723

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

Cherkashin, Ivan & Demidova, Svetlana & Imai, Susumu & Krishna, Kala, 2009. “The inside scoop: Acceptance and rejection at the journal of international economics,” Journal of International Economics, Elsevier, vol. 77(1), pages 120-132, February.

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

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

Fafchamps, Marcel & Marco J. van der Leij & Sanjeev Goyal, 2006. “Scientific Networks and Co-authorship,"Economics Series Working Papers 256, University of Oxford, Department of Economics.

Goyal, Sanjeev & Marco van der Leij & José Luis Moraga-González, 2004. “Economics: An Emerging Small World?” Tinbergen Institute Discussion Papers 04-001/1, Tinbergen Institute.

Goyal, Sanjeev, Lorenzo Ductor, Marcel Fafchamps and Marco J. van der Leij, 2011. “Social Networks and Research Output” Working paper available at http://www.econ.cam.ac.uk/faculty/goyal/wp11/prediction_july2011b%5B1%5D.pdf

Hamermesh, Daniel S., and Sharon S. Oster. 1998. “Aging and Productivity among Economists.” The Review of Economics and Statistics, 80(1): 154-156.

Hamermesh Daniel S. & Peter Schmidt, 2003. “The Determinants of Econometric Society Fellows Elections,” Econometrica, Econometric Society, vol. 71(1), pages 399-407, January.

Hoberg, Gerard and Gordon Phillips, 2010. “Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis.” Review of Financial Studies 23 (10), 3773-3811.

Laband, David N., and Michael J. Piette. 1994. “Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors.”The Journal of Political Economy, 102(1): 193-204.

Kim, E. Han, Morse, Adair and Zingales, Luigi , “Are Elite Universities Losing Their Competitive Edge? (April 2006).” CRSP Working Paper No. 609; Ross School of Business Paper No. 1046. Available at SSRN: http://ssrn.com/abstract=900920

Salton, G. & McGill, M. J., 1983. “Introduction to modern retrieval,” New York: McGraw-Hill Book Company.

Trivedi, Pravin K. 1993. “An Analysis of Publication Lags in Econometrics.” Journal of Applied Econometrics, 8(1): 93-100.