Are delays in academic publishing necessary?
Comment on Leslie (AER, 2005)
Christopher Cotton*
September 27, 2010

Abstract

Leslie (AER, 2005) shows that both submission fees and time delays at top-tier journals can maximize journal quality by discouraging long-shot submissions. He concludes that delays in academic publishing are not necessary, and argues that journals should increase submission fees in an effort to minimize delays. We show that his conclusions depend crucially on the assumption that submission costs affect all authors equally. With heterogeneous authors, maximizing journal quality requires a combination of moderate fees and moderate delays. Minimizing delay or not charging a submission fee will benefit some authors, hurt others, and strictly decrease journal quality.

Keywords: journal submission, submission fees, screening

*Department of Economics, University of Miami, Coral Gables, FL 33146; cotton@business.miami.edu.
1 Introduction

In his important contribution, Leslie (2005) recognizes that both journal submission fees and time lags (e.g., time to first response) impose costs on authors, and will discourage submissions with a low probability of acceptance. In equilibrium, submission costs screen out submissions from low-probability articles, and the editor is left with submissions that are worth her time to review. That is, fees and delays help the editor maintain an acceptable refereeing burden. In that setting, where all authors share the same costs from fees and delays, there is a continuum of fee-delay combinations that maximize journal quality. An editor is indifferent between any fee-delay combination that imposes the optimal total submission cost, whether it involves a high fee and a short delay, a low fee and a long delay (i.e., the current case at most journals), or any combination in between. When Leslie asks “Are delays in academic publishing necessary?”, he concludes that they are not. Leslie goes on to argue that the profession would benefit if top-tier journals worked to minimize the time delay associated with submission, which would require an increase in monetary submission fees. Heintzelman and Nocetti (2009) support this conclusion when they consider the effect of submission fees and delays in a similar model with multiple journals.¹ We show that Leslie’s conclusion depends crucially on the assumption that fees and delays have an identical affect on all authors.

In reality authors differ in their willingness to pay fees or deal with delays. Some authors work at institutions that pay their submission fees (for simplicity, we call these authors “rich”), others pay their fees out of pocket (i.e., “poor” authors). Similarly, untenured authors face a higher cost of delay compared to tenured authors. Our paper incorporates author preference heterogeneity in an article submission game based on Leslie (2005). When authors have heterogeneous preferences, there is a unique combination of a positive fee and time delay that maximizes journal quality. This is in contrast to Leslie, where there were many optimal combinations. Here, increasing fees in an effort to minimize delays has a disproportionately negative impact on the willingness of poor authors to submit papers and a positive impact on the willingness of rich authors to submit papers. From the perspective of journal quality, when fees are above and delays shorter than the optimal level, too many rich authors submit articles with low probability of acceptance, and too few poor authors submit articles even when they have a relatively high probability of acceptance. Similarly,

¹Azar (2005, 2007) shows that delays may be beneficial, but does not consider the tradeoff between fees and delays. Other articles consider additional aspects of the submission process. For example, Blank (1991) considers the effects of a blind review process, Engers and Gans (1998) considers why journals typically do not pay referees, and Laband and Piette (1994) look at editorial favoritism in the review process. Like Heintzelman and Nocetti (2009), Oster (1980) considers the optimal order of journals for authors submitting articles.
when fees are below and delays longer than the optimal, the journal attracts too many submissions from tenured authors and too few from untenured authors.

The optimal submission mechanism from the standpoint of journal quality imposes both moderate fees and moderate delays on authors. Relying too heavily on fees to maintain an acceptable refereeing burden makes rich authors better off, but hurts poor authors and journal quality. Relying too heavily on delays makes tenured authors better off, but hurts untenured authors and journal quality. In contrast to Leslie, we show that delays are necessary to maximize journal quality.

The model does not suggest that journals use delays in a direct effort to impose costs on authors. Rather, it suggests that journals may accept delays as a part of the review process, and that delays may increase or decrease as time constrained editors react to changes in the type and quantity of submissions. The story is consistent with Ellison (2002b,a) who argues that time-delays at top economics journals have increased over time in reaction to increased pressure to publish at these journals.

2 Model

The model is based on Leslie (2005). The primary difference between the two frameworks is that we allow for author preference heterogeneity.

There are many authors, indexed by \( i \), who may submit a paper to a top-tier journal. The continuum of authors is of total mass 1. All authors share a common value of publication, \( v > 0 \), which represents the additional benefit of publishing at a top-tier journal compared to another outlet.\(^2\) To be published, an author must first submit his paper (which may be costly), and then be selected for publication by the editor.

Papers are either high-quality or low-quality. The editor receives payoff 1 for each high-quality article published, and she receives a negative payoff for each low-quality article published. Author \( i \)'s paper is high-quality with probability \( p_i \). Authors privately observe their own \( p \), but not whether their paper is high or low quality. Each \( p_i \) is the independent realization of a random variable uniformly distributed on \([0, 1]\). The editor knows the distribution of \( p \), but does not observe the draws.\(^3\)

If the editor reviews a submission, she becomes certain of its quality. However, the editor is time constrained and may only review submissions from up to \( \tau < 1 \) portion of the

\(^2\)Leslie (2005) uses \( V(1) \) and \( V(2) \) to denote the value of a top-tier publication and the value of the outside option. Thus, \( v = V(1) - V(2) \).

\(^3\)This discrete framework captures all of the necessary aspects of Leslie (2005)'s continuous-quality model. In both frameworks, the editor accepts a paper only if it is “high-enough” quality, and authors privately observe the probability that the article is high-enough quality.
After reviewing a paper, the editor decides whether to publish it. She cannot publish an article she does not review. Since the editor fully learns a submission’s type before acceptance, she always accepts high-quality submissions and never accepts low-quality submissions.

The editor chooses fee $m \geq 0$ and time requirement $t \geq 0$ to impose on reviewed submissions. Author $i$ experiences costs $-\alpha_i m - \beta_i t$ from submission, independent of acceptance.

We assume $\alpha_i, \beta_i \in \{0, 1\}$. That is, some authors (whom we refer to as “untenured”) experience disutility from time delays, while other others (i.e., “tenured”) do not. Similarly, some authors (e.g., “poor” authors who pay fees out of pocket) experience disutility from submission fees, while others (e.g., “rich” authors with generous research budgets) do not. An author has $\alpha_i = 1$ with probability $\pi_m \in (0, 1]$, and $\beta_i = 1$ with probability $\pi_t \in (0, 1]$. The realizations of $\alpha_i$ and $\beta_i$ are independent. All author types share the same value of publication, $v$. Allowing authors to also differ in terms of $v$ is realistic, but complicates the analysis without changing the main result. It would still be the case that journal quality is maximized through a combination of fees and delays, and that setting too high of fees or too long of delays will benefit some authors, hurt other authors, and decrease journal quality.

The editor wants to maximize journal quality, which we define as the total number of high-quality papers published. Let $a_i \in \{0, 1\}$ indicate whether article $i$ is high quality and accepted. The editor receives total utility equal to the average value of $a_i$. Author $i$’s payoffs are $u_i = a_i v - \alpha_i m - \beta_i t$ if he submits his article for review, and $u_i = 0$ if he does not submit his article.

The editor does not directly benefit from collecting fees. This is a simplifying assumption that serves to focus the analysis on the ability of costs to screen submissions. If the editor can use the submission fees to pay referees or pay for operation costs, the use of fees will become relatively more attractive for the editor. The main results, however, should not change: relying too heavily on fees and not enough on delays will decrease journal quality.

The editor must review all submissions. If $\lambda(m, t)$ denotes the portion of authors that submit papers given $m$ and $t$, then the editor is constrained to choosing $(m, t)$ such that

---

4 This is similar to the acceptable refereeing burden in Leslie (2005). The assumption that the editor fully learns an article’s type upon review is also taken from Leslie, and is equivalent to the case of perfect referee reports in Heintzelman and Nocetti (2009).

5 The time delay $t$ represents the expected delay from submission. We ignore the fact that a minimum amount of time delay will be required to conduct a review. Allowing for a minimum amount of delay will not change any of the results, assuming that the minimum delay amount is sufficiently low.

6 Such assumptions about the distribution of $\alpha_i$ and $\beta_i$ simplify the analysis, but are not necessary for the results. Not allowing $\pi = 0$ rules out the possibility that no authors are affected by fees or delays.

7 Furthermore, as Leslie (2005) points out, if untenured authors have a higher $v$ than tenured authors, the optimal cost mechanism may account for this by changing higher submission fees to untenured authors.

8 For a discussion on paying referees, see Engers and Gans (1998) and Chang and Lai (2001).
\( \lambda(m, t) \leq \tau \). To ensure that this is always feasible, even in the cases where the editor is constrained to set \( m = 0 \) or \( t = 0 \), assume \( \tau > \max\{1 - \pi_m, 1 - \pi_t\} \); otherwise there may be more submissions than time, independent of the editors choice of cost mechanism. The results are unchanged if we allow the editor to set \( (m, t) \) such that \( \lambda > \tau \) and then randomly select \( \tau / \lambda \) of those submissions to review, as long as \( \tau \) is not too small. This restriction on \( \tau \) allows us to focus on developing intuition, not solving for corner solutions; it is not necessary for our results. Finally, an author who is indifferent about submitting and not submitting will submit.

The game takes place as follows:

1. The editor chooses the submission mechanism, which consists of \( m \) and \( t \).
2. Each author chooses whether to pay \( m \) and \( t \) to submit his paper.
3. The editor reviews submissions, then chooses which articles to publish.

## 3 Equilibrium

The analysis solves for the subgame perfect Nash equilibrium. We solve for equilibrium using backward induction.

As we discuss in the previous section, in the final stage of the game the editor will always publish high-quality submissions, and never publish low-quality submissions. Given this, authors submit their papers when their signals are high enough. That is, \( i \) submits if and only if his quality signal \( p_i \) is at least as great as cutvalue \( \bar{p}_i \), where

\[
\bar{p}_i \equiv \frac{1}{v} (\alpha_i m + \beta_i t).
\]

When \( p_i \geq \bar{p}_i \), the author’s expected payoff from submitting is positive. Eq. 3 fully describes the equilibrium strategy for the authors. Next, we determine the editor’s choice of \( m \) and \( t \).

The editor can limit the number and expected quality of submissions by imposing costs on authors. When there are no costs (i.e., \( m = t = 0 \)), all authors submit their articles. Fees impose costs on and reduce submissions from authors with \( \alpha_i = 1 \). Similarly, delays impose costs on and limit submissions from authors with \( \beta_i = 1 \). An author with \( \alpha_i = \beta_i = 1 \) is affected by both types of costs, while an author with \( \alpha_i = \beta_i = 0 \) is unaffected by costs and will always submit his paper even when costs are high. The optimal mix of \( m \) and \( t \) depends on how valuable publication is the the authors, and the portion of authors of each type.

Let \( m^* \) and \( t^* \) denote the optimal choice of \( m \) and \( t \) to maximize journal quality.
Proposition 1  Journal quality is maximized when

\[ m^* = \frac{b(1-\pi_t)(1-\tau)}{\pi_t(1-\pi_m)+\pi_m(1-\pi_t)} \quad \text{and} \quad t^* = \frac{b(1-\pi_m)(1-\tau)}{\pi_t(1-\pi_m)+\pi_m(1-\pi_t)}. \]

This is a unique maximum. Any \((m, t) \neq (m^*, t^*)\) decreases journal quality.

Corollary 2 follows immediately from Proposition 1. It shows that when authors are heterogeneous on both cost dimensions, the optimal cost mechanism involves both positive fees and positive delays.

Corollary 2

- \(m^* > 0\) and \(t^* > 0\) when \(\pi_m, \pi_t \in (0, 1)\),
- \(m^* = 0\) only if \(\pi_t = 1\), and
- \(t^* = 0\) only if \(\pi_m = 1\).

When authors differ in both \(\alpha\) and \(\beta\), it is optimal for the editor to use both fees and delays. Doing so levies more-uniform total costs across the entire author population. If she only relies on delays and not fees, the tenured-author cutvalue, \(\bar{p}\), will be too low and the untenured-author signal cutvalue will be too high. From the standpoint of journal quality, this means too many tenured authors and too few untenured authors submit papers. If she only relies on fees and not delays, the rich-author cutvalue will be too low and the poor-author cutvalue too high. From the standpoint of journal quality, too many rich authors and too few poor authors submit papers.

Imposing only one type of cost is optimal only if the authors are homogeneous along that cost dimension. For example, if all authors share the same disutility from delays (i.e., \(\pi_t = 1\)), then the editor can impose the same submission costs on all authors by relying only on delays. This means that not charging submission fees is only optimal if all authors find delays equally costly. Given the tenure clock, this is unlikely. A journal that does not charge a submission fee will benefit from implementing one. Similarly, when all authors find monetary payments equally costly (i.e., \(\pi_m = 1\)), it is optimal to rely only on fees.

Leslie (2005) and Heintzelman and Nocetti (2009) argue that fees should be increased in order to minimize delay. In their framework with homogeneous authors, there exists a continuum of fee-delay combinations which all result in maximum journal quality. This means that the editor can increase fees and decrease delays without adversely affecting quality. In our model, where authors differ in their cost preferences, there is a unique combination of fees and delays which maximizes journal quality. In this case, minimizing
delay, or not charging a submission fee, decreases equilibrium journal quality. Proposition 3 considers the impact that such changes in $m$ and $t$ have on average payoffs.

**Proposition 3** Let $\pi_m \in (0, 1)$ and $\pi_t \in (0, 1)$. Compared to the optimal $m^* > 0$ and $t^* > 0$,

- minimizing delays (i.e., choosing $m$ to maximize journal quality, subject to $t = 0$), increases the average payoff of rich authors, and decreases both the average payoff of poor authors and journal quality; and

- not charging a submission fee (i.e., choosing $t$ to maximize journal quality, subject to $m = 0$), increases the average payoff of tenured authors, and decreases both the average payoff of untenured authors and journal quality.

4 Discussion

This paper presents a simple model of the academic publication process in which authors may face submission fees and time delays from submitting a paper. Both fees and delays help reduce submissions from authors with only a “long-shot” of acceptance. This helps journals maintain an acceptable refereeing load, and ensures that editors use their limited time to review the most-promising articles.

Leslie (2005) considers a similar settings, but assumes that fees and delays affect all authors equally. When authors are homogeneous, there are many combinations of fees and delays that may be imposed to effectively screen out low-quality authors. Given this, Leslie argues that it is best to impose high fees and minimize delays. Doing so minimizes the effective time to publication, and can encourage the quick dissemination of research.

In our paper, authors are heterogeneous, in that not all authors are equally affected by time delays and fees. In contrast to the case when authors are homogeneous, here increasing fees in order to minimize delay results in too many submissions from authors who are unconcerned about paying the fees, and too few submissions from authors with less generous research budgets who must pay fees out of pocket. This increases the payoffs of “rich” authors, but decreases “poor” author payoffs and journal quality. If, on the other hand, the editor imposes no fee and instead relies only on delay to limit the number of submissions, then too many tenured authors submit papers, and too few untenured authors submit papers. This increases average tenured author payoffs, but decreases untenured author payoffs and journal quality.

Leslie (2005) noted in his concluding discussion that submission costs will to account for author heterogeneity, suggesting that universities not pay their faculty’s submission fees.
Even if eliminating universities subsidization of submission fees was feasible, these recommendations would hardly eliminate all differences in willingness to pay fees among authors, which would be required for his recommendations to not decrease journal quality. Some authors will always have higher income or more generous research budgets than other authors, and these authors will submit too many papers and the other authors will submit too few papers if the journal relies only on fees.

Our results highlight a downside of journals relying only on delays or only on fees to maintain an acceptable refereeing load. The analysis supports Leslie (2005)’s recommendation that journals move away from a system of high delays and low (if any) monetary submission fees. However, we also show that it would be a mistake to over do it, which would likely be the case if journals followed Leslie’s recommendation and moved to a system of high fees and minimum feasible delay. Relying too heavily on fees results in many of the same inefficiencies as relying too little on fees. A combination of moderate fees and delays provides the most efficient screening of submissions and maximizes overall journal quality.

5 Appendix

Proof of Proposition 1. Given $m$ and $t$, the portion of authors who submit papers simplifies to

$$\lambda = 1 - \frac{\pi_mm + \pi_tt}{v}.$$ 

The editor chooses $m$ and $t$ to maximize journal quality,

$$\frac{\tau}{\lambda} \left[ \pi_mm\int_{v}^{1} pdp + \pi_m(1 - \pi_t)\int_{\frac{1}{v}}^{1} pdp + (1 - \pi_m)\pi_t\int_{v}^{1} pdp + (1 - \pi_m)(1 - \pi_t)\int_{0}^{1} pdp \right],$$

subject to $0 \leq m$, $0 \leq t$, and $\tau \geq \lambda$. The editor will never choose $m$ and $t$ such that $\lambda < \tau$, since should could reduce $m$ or $t$ and review additional articles. Thus $\tau = \lambda$. Maximizing journal quality with respect to $m$ and $t$, subject to $\tau = \lambda$, $0 \leq m$, and $0 \leq t$ yields the equilibrium solution, denoted $(m^*, t^*)$, which is provided in the proposition. A check of the second order conditions ensures this is a maximum. Uniqueness of equilibrium follows from the single solution to the editor’s maximization problem.

Proof of Corollary 2. Consider the equilibrium expressions for $m^*$ and $t^*$. If $\pi_m \in (0, 1)$ and $\pi_t \in (0, 1)$, then both $m^* > 0$ and $t^* > 0$. $m^* = 0$ only if $\pi_t = 1$, and $t^* = 0$ only if $\pi_m = 1$.

Proof of Proposition 3. Under $m^*$ and $t^*$, the poor-untenured author cutvalue is $\bar{p}_{PU} = \frac{(1-\tau)(1-\pi_t)(1-\pi_t)}{\pi_m(1-\pi_t)+\pi_t(1-\pi_m)}$, and the rich-untenured cutvalue is $\bar{p}_{RU} = \frac{(1-\tau)(1-\pi_t)}{\pi_m(1-\pi_t)+\pi_t(1-\pi_m)}$. An
author's cost of submitting equals his $p_i$ times $v$. Thus, the average payoff to untenured authors equals
\[ \pi_m \int_{\bar{p}_{PU}}^{1} vpd\bar{p} + (1 - \pi_m) \int_{\bar{p}_{RU}}^{1} vpd\bar{p} - (\pi_m\bar{p}_{PU}v + (1 - \pi_m)\bar{p}_{RU}v). \]

Similarly, the poor-tenured author cut value equals $\bar{p}_{PT} = \frac{(1-\tau)(1-\pi_t)}{\pi_m(1-\pi_t)+\pi_t(1-\pi_m)}$, and the rich-tenured cut value equals $\bar{p}_{RT} = 0$. Thus, the average payoff to tenured authors equals
\[ \pi_m \int_{\bar{p}_{PT}}^{1} vpd\bar{p} + (1 - \pi_m) \int_{0}^{1} vpd\bar{p} - \pi_m\bar{p}_{PT}v. \]

It is straightforward to solve the editor's optimization problem for $t > 0$, subject to $m = 0$. Denote this optimal choice of $t$ given $m = 0$ by $\hat{t}$. When $m = 0$, $\hat{t} = \frac{\pi(1-\tau)}{\pi_t}$. In this setting, with $m = 0$ and $\hat{t} > 0$, the untenured author cut value equals $\bar{p}_{U} = \frac{1-\tau}{\pi_t}$ for both rich and poor; thus, $\bar{p}_{U} = \bar{p}_{PU} = \bar{p}_{RU}$. Similarly, the tenured author cut value equals $\bar{p}_{T} = 0$ for both rich and poor. The average payoffs may be calculated in the same way as before.

By comparing these average expected payoffs for $(m, t) = (m^*, t^*)$ and $(m, t) = (0, \hat{t})$, we see that (1) untenured authors are strictly better off under $(m^*, t^*)$, and (2) tenured authors are strictly better off under $(0, \hat{t})$. When the editor maximizes journal quality, defined as the number of high-quality articles published, she chooses $m^* > 0$ and $t^* > 0$ from the proof to Prop 1. This is true even though she was permitted to choose set $m = 0$. Thus, maximum journal quality is achieved at $m^* > 0$ and $t^* > 0$, not $m = 0$ and $t = 0$.

A symmetric analysis may be conducted for rich and poor candidates, under $m^*$ and $t^*$ and for the case where the editor chooses $\hat{m} > 0$ subject to $t = 0$. By comparing these average expected payoffs for $(m, t) = (m^*, t^*)$ and $(m, t) = (\hat{m}, 0)$, we can show that (1) poor authors are strictly better off under $(m^*, t^*)$, (2) rich authors are strictly better off under $(\hat{m}, 0)$, and (3) journal quality is maximized under $(m^*, t^*)$.\[\square\]

References


