NBER WORKING PAPER SERIES

CERTIFIED RANDOM:
A NEW ORDER FOR CO-AUTHORSHIP

Debraj Ray ⓡ
Arthur Robson

Working Paper 22602
http://www.nber.org/papers/w22602

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2016

Ray thanks the National Science Foundation for support under grant numbers SES-1261560 and SES-1629370. Robson thanks the Canada Research Chairs Program and the Social Sciences and Humanities Research Council of Canada. We thank four referees, Nageeb Ali, Dan Ariely, Joan Esteban, Itzhak Gilboa, Ed Green, Johannes Horner, Navin Kartik, Laurent Mathevet, Sahar Parsa, James Poterba, Andy Postlewaite, Phil Reny, Ariel Rubinstein, Larry Samuelson, Rakesh Vohra, and Leeat Yariv for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Certified Random: A New Order for Co-Authorship
Debraj Ray ⓡ Arthur Robson
NBER Working Paper No. 22602
September 2016
JEL No. A10,A14

ABSTRACT

Alphabetical name order is the norm for joint publications in economics. However, alphabetical order confers greater benefits on the first author. In a two-author model, we introduce and study certified random order: the uniform randomization of names made universally known by a commonly understood symbol. Certified random order (a) distributes the gain from first authorship evenly over the alphabet,(b) allows either author to signal when contributions are extremely unequal, (c) will invade an environment where alphabetical order is dominant, (d) is robust to deviations, (e) may be ex-ante more efficient than alphabetical order, and (f) is no more complex than the existing alphabetical system modified by occasional reversal of name order.

Debraj Ray
Department of Economics
New York University
19 West Fourth Street
New York, NY 10003
and University of Warwick
and also NBER
debraj.ray@nyu.edu

Arthur Robson
Department of Economics, Simon Fraser University
8888 University Drive
Burnaby, British Columbia, CANADA V5A 1S6
robson@sfu.ca

1. BACKGROUND

Our last names above appear in alphabetical order, but a coin was tossed to determine name placement. The symbol ⓡ between our names is a signal that the names are in *random order*. *Certified* random order — randomization that is institutionally marked by a commonly understood symbol such as ⓡ — is the topic of this paper.

Alphabetical order is the norm for name order in joint research in economics. Table 1 reports the prevalence of this norm. Around 85% of two-author economics papers are written with the authors listed in alphabetical order.[2] That percentage falls with more authors, possibly capturing the fear of *et al* oblivion, or there could be research assistants involved.[3] Compare this to the physical sciences, in which first authorship is given — presumably not without occasional disagreement — to the lead contributor, while other not-so-subtle signals such as lab leadership are sent through ancillary ordering conventions. Possibly the civility of the alphabetical norm lends itself to more joint work, as the possible rancor in settling on a name order at publication time is thereby avoided.

And yet, there are serious issues with alphabetical order:

1. Psychologically, names that appear first are more likely to be given "extra credit." This order effect is certainly in line with research on marketing: products presented earlier exhibit higher probabilities of selection, as the aptly ordered article by Carney and Banaji (2012) observes. Even stocks with earlier names in the alphabet are more likely to be traded; see another aptly ordered paper by Jacobs and Hillert (2016), or Itzkowitz, Itzkowitz and Rothbort (2016).

2. Earlier names appear bunched together on a bibliographical or reference list, promoting the citation of the paper. Haque and Ginsparg (2009) — aptly ordered again — note that article positioning in the ArXiv repository is correlated with citations of that article. Feenberg, Ganguli, Gaule, and Gruber (2015) demonstrate that the same bias exists in the downloading and citation of NBER "New This Week" Working Papers, which led to a change in NBER Policy.[4]

[2]Certainly, alphabetical order is occasionally overturned (see Table 1 again) and when it is, it is a clear signal that the author who now appears first has done the bulk of the work. This option is central to the theory we develop.

[3]The influence of the *et al* possibility is possibly captured better by papers in which only the first author is out of alphabetical order; this percentage is, inevitably, lower as Table 1 reveals.

[4]An email from James Poterba dated September 2, 2015, states that "beginning next week, the order of papers in each of the more than 23,000 "New This Week" messages that we send will be determined randomly. This will mean that roughly the same number of message recipients will see a given paper in the first position, in the second position, and so on."

	Number of Authors			
	Two	**Three**	**Four**	**Five**
Total	53858	17857	1865	340
Alphabetical	45337	13124	1155	163
Non-Alphabetical	8521	4733	710	177
% Non-Alpha	**15.82**	**26.51**	**38.07**	**52.06**
First Author Non-Alpha	8521	2754	339	95
% First Author Non-Alpha	**15.82**	**15.42**	**18.18**	**27.94**

Table 1. Alphabetical Order in Peer-Reviewed Journals in Economics. *Sources and Notes. EconLit*, 1969–2013, using the list of 69 leading economics journals in Engemann and Wall (2009).

3. There is at least one major journal in economics (the *Review of Economic Studies*) which publishes articles in alphabetical order (using the last name of the first author). Because many other journals use the convention that the lead article is special, and because many do not know that the *Review of Economic Studies* follows this policy, this confers a potential advantage on earlier names.

4. The *et al* convention, which is widely used in citations and especially on slides in seminars, obscures the identity of later authors. Even if *et al* were to be banned in journal publications, it cannot be banned from slides. In addition, it is widespread practice in verbal presentations to mention the name of the first author and then add "and coauthors": an understandable but inequitable shortcut.

There is good evidence that these considerations matter. In a paper published in the *Journal of Economic Perspectives*, Einav and Yariv (2006) write (we quote their abstract in full):

> "We present evidence that a variety of proxies for success in the U.S. economics labor market (tenure at highly ranked schools, fellowship in the Econometric Society, and to a lesser extent, Nobel Prize and Clark Medal winnings) are correlated with surname initials, favoring economists with surname initials earlier in the alphabet. These patterns persist even when controlling for country of origin, ethnicity, and religion. We suspect that these effects are related to the existing norm in economics prescribing alphabetical ordering of authors' credits. Indeed, there is no significant correlation between surname initials and tenure at departments of

> psychology, where authors are credited roughly according to their intellectual contribution. The economics market participants seem to react to this phenomenon. Analyzing publications in the top economics journals since 1980, we note two consistent patterns: authors with higher surname initials are significantly less likely to participate in projects with more than three authors and significantly more likely to write papers in which the order of credits is non-alphabetical."

There are other papers that buttress the Einav-Yariv empirical findings; see, for instance, the impeccably hedged Chambers, Boath and Chambers (2001), or the unavoidably unordered van Praag and van Praag (2008). Going beyond Einav and Yariv (2006), this last article finds "significant effects of the alphabetic rank of an economist's last name on scientific production, given that an author has already a certain visibility in academia ... Being an A author and thereby often the first author is beneficial for someone's reputation and academic performance." Moreover, as they go on to observe, the recognition accorded to earlier authors appears to cumulate over time: "Professor A, who has been a first author more often than Professor Z, will have published more articles and experienced a faster productivity rate over the course of her career as a result of reputation and visibility." A recent survey by Weber (2016) summarizes the literature thus: "there is convincing evidence that alphabetical discrimination exists."

2. Name-Order Conventions and Institutions

Social conventions — name-order norms here — may be viewed as equilibria that are immune to deviations by individuals (or segments of the population). For instance, can private randomization arise as a deviation from alphabetical order? Suppose that Jane Austen and Lord Byron, working together,[5] contemplate the randomization of their joint authorship, perhaps over a sequence of papers. There are difficulties. *Given* an "alphabetical society," an order reversal is a clear signal that the newly christened first author has done the bulk of the work. That is, "Byron and Austen" would be a statement that Byron has done most of the research on the paper, whereas "Austen and Byron" would indicate very little, any such signal being swallowed in most part by the naming convention. Therefore Austen gains nothing over alphabetical order when her name comes first,

[5]This stretches realism a bit. Although Austen and Byron were contemporaries, there is no evidence of their meeting, let alone collaborating. A key advantage of this pairing of coauthors is that it makes the use of "he" or "she" unambiguous.

while Byron gains a lot when his does.[6] Byron will agree to the *ex-ante* randomization, but Austen will not. This is true even when their randomization is recorded in the publication itself; say in a lead footnote. After all, research often becomes visible through written citations and verbal references rather than direct persual; see Section 7.1 for more discussion. We will formalize these remarks by showing that alphabetical order is robust to deviations — deterministic or random — given the set of alternatives available to authors today (Theorem 1).

But institutions can change that. Here is a simple variant of the randomization scheme which will set it apart from private randomization. Suppose that any randomized name order is presented with the symbol ⓡ between the names; e.g., Ray ⓡ Robson (2016) is the appropriate reference for this paper. Suppose, moreover, that such a symbol is certified by the *American Economic Association*, for example, simply acknowledging that this alternative is available.

It is unclear that this "mutant" would successfully take over the population. But we are going to argue that it will. The key point that makes this argument possible is that economics does not *entirely* follow alphabetical order. There are exceptions, which occur when the author who is lower down the alphabetical food chain has really contributed disproportionately. These exceptions are made quite often. Table 1 shows that over 15% of two-author publications in the 69 leading economics journals identified by Engemann and Waall (2009) have their names reversed. That percentage rises significantly for three or more authors.

How are these exceptions made? Presumably the first author concedes the reversal in circumstances where the second author has made a much larger contribution. The exact source of the concession is unimportant. It may be a sense of fair play or guilt. It may be an aversion to an unpleasant conversation. It might be the result of a rational calculation made by the first author when agreeing or disagreeing to reverse name order. For instance, the second author might refuse to collaborate again with the first author if he feels he has been treated unfairly. We summarize

[6]Engers *et al* (1999) emphasize this point, arguing that alphabetical order can *disadvantage* "early authors," because a reversal can be used to signal a higher contribution by the late author, but there is no comparable signal for the early author. That may well be true, but on the other hand Engers *et al* have no counterpart to the direct premium from first-authorship that we will posit. The empirical literature that we've discussed suggests that such a premium is a first-order consideration, and indeed it is the central motivation for our paper. The Engers *et al* model does not generate the advantage to first authorship seen in the data, because the authors' payoffs are only the Bayesian rational assignment of credit. For example, if authors are always listed alphabetically, then the credit assigned in their model will be equal.

all this by introducing a suitable loss function that is experienced by the first author if the second author has a high contribution relative to the credit he publicly receives.

It will turn out that this capacity for name reversal also facilitates certified random order, but not private randomization. The key reason that ⓡ outperforms private randomization is that the notation ⓡ is permanently attached to all subsequent references to the paper, whereas private randomization is not.[7] Once certified random order is introduced, it breaks the alphabetical order equilibrium, even though this equilibrium is robust to the possibility of private randomization (Theorem 2). The reverse is not true: random order is robust to invasion by alphabetical order or by reverse order (Theorems 3 and 4).

The possibility of successful invasion does *not* imply that the replacement is "socially better." There is a constant-sum component of the payoffs— if, for example, Austen achieves a higher credit based on the social perception of her contribution, that necessarily reduces the social credit share for Byron. However, the cost that is generated by a gap between an author's actual contribution and that imputed to him or her by society introduces a nonzero sum component to the payoffs. The two conventions considered here can therefore differ in terms of the sum of expected payoffs. Theorem 5 illustrates this by comparing the new ⓡ convention and the current convention for the tractable special case in which the distribution of contributions is uniform. Here, ⓡ achieves a higher sum of expected payoffs than the Economics convention.[8] In this example, then, any quasiconcave Bergson-Samuelson welfare function defined over author payoffs would prefer random order to alphabetical order.

To summarize, ⓡ can be introduced not as a requirement but as a nudge, because our results predict that it will invade alphabetical order in a decentralized way. It may provide a gain in efficiency. But, more importantly, it is fairer. Random order distributes the gain from first authorship evenly over the alphabet. Moreover, it allows "outlier contributions" to be recognized in both directions;

[7]Even if a lead footnote details the randomization, this information is lacking in subsequent references.

[8]There are other efficiency arguments. For instance, individuals put effort into doing research. Unequal division of the credits from that research might be surplus-dominated — even Pareto-dominated — by equal division, as efforts adjust to the more equitable distribution of credits. This approach to team production with moral hazard is not followed here. Engers *et al* (1999) derive the contributions of authors from endogenous effort choices. They show that alphabetical order prevails over meritocracy in equilibrium, despite the greater efficiency of the latter (in their model). There are also possible efficiency losses from the strategic choice of co-authors when it is feared that lexicographic relegation to the end of the name order (or even to the anonymity of *et al*) might lead to a decreased payoff. Einav and Yariv (2004) provide evidence that individuals further down the alphabet are more averse to writing with multiple co-authors. See Ray (2013) for notes on strategic choice of co-authors.

that is, given the convention that puts "Austen ⓡ Byron" (or "Byron ⓡ Austen") on center-stage, *both* "Austen and Byron" and "Byron and Austen" would acquire entirely symmetric meanings. Finally, except for the addition of a simple symbol, random order is no more complex than the existing alphabetical convention.

Now for the details.

3. A Model of Name-Order Conventions

Suppose a paper is worth a total credit of 1 unit. There are two authors: Austen and Byron.[9] Their contributions are x and $1-x$ respectively, where x is distributed on $[0,1]$ with strictly positive density given by f. Ex post, $(x, 1-x)$ is observed by the authors but not by the public, who must infer these shares from the social convention in force and the particular name order followed.

We allow for a general class of distributions, including those that are asymmetric. Asymmetry may stem from co-author characteristics that are publicly observable, such as professor-student pairs, or the presence of a particularly eminent co-author.

3.1. Conventions and Defaults.

Let n be a naming scheme—that is, alphabetical order ($n = \alpha$), reverse-alphabetical order ($n = \rho$), or certified random order ($n = \textcircled{r}$). We assume that, in each naming scheme, names must be presented sequentially.[10] There is some set of available naming schemes. For the Economics convention, this set is formally $\{\alpha, \rho\}$; for the certified random order convention, it is enlarged to $\{\alpha, \rho\, \textcircled{r}\}$. We also permit private randomization across naming schemes.

The use of a particular scheme sends signals about the contributions $(x, 1-x)$. But these signals also depend on the naming *convention* in place in society: a map from contributions $(x, 1-x)$ to allowable naming schemes. For instance, *pure meritocracy* is the convention that chooses α when $x > 1/2$, and ρ when $x < 1/2$. *Pure alphabetical order* is the convention that chooses α no matter what the contributions are.

A convention is associated with a *default action*, one that either party can insist on. The Economics convention is close to pure alphabetical order, with the occasional reversal of name order to signal a

[9] The analysis of three or more authors is an interesting open question.

[10] That is, in every scheme, one name is stated first, then the other. This is certainly true of any spoken scheme.

significant imbalance in contributions. We therefore suppose that the default under the Economics convention is alphabetical order, which is invoked in the event of disagreement, and can be insisted upon by either party.

3.2. **Credit.** Let $A(n, C)$ and $B(n, C)$ denote the socially imputed credits to Austen and Byron respectively under convention C when the naming scheme n is followed. These social credits are respectively the *conditional expectations* of x and $1 - x$ entailed by n under the convention C. (These are not overall payoffs, which will include two other components to be described.) Economics uses what might be described as a modified alphabetical convention E_t with name reversal when Austen's contribution drops below some threshold t. Then the use of $n = \alpha$ yields a credit of

$$A(\alpha, E_t) = h(t) \text{ and } B(\alpha, E_t) = 1 - h(t) \tag{3.1}$$

to Austen and Byron respectively, where $h(t)$ stands for the expectation of Austen's contribution conditional on that contribution exceeding t. Likewise, if ρ is observed, the corresponding credits are

$$A(\rho, E_t) = l(t) \text{ and } B(\rho, E_t) = 1 - l(t), \tag{3.2}$$

where $l(t)$ is the expectation of Austen's contribution conditional on that contribution falling short of t.[11]

Of course, $A(n, C) + B(n, C) = 1$; that is, a total credit of 1 is always being divided.[12]

3.3. **Reputational Payoff.** We suppose that each name order yields a gain $\delta > 0$ in the reputation of the *first* author in the order. This is due to visibility, bunching in reference lists, the *et al* effect and so on, and it accrues over and above the "direct credits" $A(n, C)$ and $B(n, C)$ that the public will estimate. This reputational payoff is central to our model.

[11] By convention, $l(0) = 0$ and $h(1) = 1$, while $h(0) = l(1)$ is the unconditional mean of Austen's contribution.

[12] It is possible that name order also affects total credit. For instance, "Byron and Austen" will garner less *overall* attention than "Austen and Byron" in a reference list. As Nageeb Ali pointed out to us, that could tilt both Austen and Byron towards "Austen ⓡ Byron" over "Byron ⓡ Austen", so that, conceivably, Austen and Byron might add ⓡ without actually randomizing. Whether Byron agrees to such a ploy depends on whether the "bibliography effect" dominates the even shot at having his name first. We are therefore assuming that the former effect is small relative to name arrangement. This issue could be addressed by requiring that all randomizations be carried out by the publishing outlet.

3.4. **Loss Function.** Our final component of the payoffs incorporates the disutility or loss generated when the imputed credit departs substantially from relative contributions. We summarize this loss by a function Γ, which is experienced by the author if her co-author has been treated badly. It has as its argument the *difference* z between the true credit due to the co-author, and the inferred credit that the co-author obtains in the public eye from the announced name order and the going convention.[13]

For example, consider the Economics convention with threshold t. Suppose that $(x, 1-x)$ are the true contributions, whereas the inferred contribution from alphabetical order under E_t is $\{h(t), 1-h(t)\}$. Then the shortfall for Byron is $z_B = [1-x] - [1-h(t)] = h(t) - x$, and the resulting loss that Austen experiences is $\Gamma(h(t) - x)$.

We impose the following restrictions on Γ. First, $\Gamma(z) = 0$ when $z \leq 0$. Second, Γ is continuously differentiable everywhere, strictly increasing and strictly convex for $z \geq 0$.[14] Finally, we impose two conditions that concern extreme outcomes. Let m_A and m_B be the unconditional expected contributions of Austen and Byron—that is, of x and $1-x$, respectively. We assume that

$$\Gamma(m_A) > m_A + \delta \text{ and } \Gamma(m_B) > m_B + \delta, \tag{3.3}$$

and

$$\Gamma\left(\frac{1}{2}\right) \leq 1. \tag{3.4}$$

Condition (3.3) can be interpreted as follows. Suppose, for example, that Byron has done *all* the work, and Austen none, so that $(x, 1-x) = (0, 1)$. Austen is then offered a binary choice between conceding full credit to Byron, thereby obtaining a payoff of 0, or taking the net payoff from alphabetical order evaluated at her expected contribution, which is $m_A + \delta - \Gamma(m_A)$. Inequality (3.3) states that Austen will wish to reverse authorship in this case. (A similar argument for Byron

[13] As mentioned above, there are multiple potential sources of such a loss. It could arise from a sense of fair-play, for example, or it could instead serve as reduced-form expression for the future consequences of short-changing a co-author.

[14] The strict convexity of Γ means that each author is increasingly intolerant of a greater divergence between the actual and imputed contribution of a coauthor. This implies that an author wants to concede first authorship to a co-author if the contribution of the latter exceeds a certain threshold, but not otherwise. See (4.3) below and the analysis around it. It might well be possible to require only that Γ be convex beyond a certain point.

yields the second inequality in (3.3).) Without some such limitation on Γ, the co-existence of alphabetical order and reversal, as in Table 1, would not be observed.

To understand (3.4), suppose that each author makes an equal contribution, so that $x = 1/2$, and that there is no δ-premium to name order. Now Austen is offered the purely hypothetical choice between being assigned full social credit for herself, with payoff $1 - \Gamma(1/2)$, or conceding full credit to Byron, which yields her 0. Inequality (3.4) asserts that Austen would then weakly prefer her name to go first, rather than reverse names. This assumption serves to limit the impact of the loss term.

3.5. **Overall Payoffs from a Convention.** Consider any convention C that maps realized contributions x to a naming scheme n that will be used in publication. Austen's overall payoffs at x are then $u_A(C, x, n)$, which is

(i) her socially imputed credit from n, which is $A(n, C)$ *plus*

(ii) δ if her name comes first under n, *or* 0 if it comes second; *minus*

(iii) the loss $\Gamma(A(n, C) - x)$ generated by n at x, as described above.

A parallel formulation holds for Byron's overall payoff $u_B(C, x, n)$. Note that a convention could also include randomizations over name order for some realizations of x, in which case we take expected values over the above payoffs.

3.6. **Equilibrium Conventions.** We now describe an *equilibrium convention*. We do not need to specify the strategic interaction between the authors explicitly. We only require that it have the following properties. Consider any $x \in [0, 1]$ where $C(x) = n$, say. Then:

[I] It cannot be that *either* author strictly prefers d to n, where d is the default naming scheme under C. That is, if $u_i(C, x, d) > u_i(C, x, n)$ for either $i = A$ or $i = B$, then $C(x) \neq n$. This requirement formalizes the favored role of the default d. In particular, given x, if for every allowable non-default naming scheme, n', some author strictly prefers d to n', then d is the only equilibrium naming scheme.

[II] It cannot be that *both* authors strictly prefer a scheme n' to the implemented n. That is, if $u_i(C, x, n') > u_i(C, x, n)$ for both $i = A$ and $i = B$, then $C(x) \neq n$.

An *equilibrium convention* is a convention satisfying [I] and [II] for every $x \in [0, 1]$. The existence of such an equilibrium convention is established by construction in each of the cases examined below. Here is a specific non-cooperative game with outcomes that satisfy [I] and [II].

Contributions $(x, 1 - x)$ are revealed, and then Austen and Byron simultaneously propose an outcome from the set of naming schemes available in the convention. If the same action is chosen by both authors, this is implemented. If the two authors take different *non*-default actions, then the default is implemented. Finally, if one author chooses the default, and the other another action n, then the author who chose the default naming scheme as the action is given the opportunity to agree to n, or make a new proposal n' of her own. If the counterproposal is accepted, it is implemented. If not, the default is implemented.

3.7. **Rational Disruption of a Convention.** Suppose a new action is added to the set of allowable actions in an equilibrium convention. For instance, suppose that ⓡ, which does not exist in the Economics convention, makes an appearance. We will model the payoffs from the use of ⓡ as arising from an accurate social perception of author contributions when this new action is taken.

More precisely, suppose that ⓡ is available only to a vanishingly small group of coauthors, so that the social assessment of all existing actions is undisturbed. This vanishingly small group has all types $(x, 1 - x)$ in it, with the same distribution as that in the population at large. On seeing ⓡ, suppose social credits of $(a^*, 1 - a^*)$ are assigned. With credits assigned to all outcomes, suppose that there is a non-negligible set P of types (within the negligible "mutant" group of coauthors) that choose the new action in equilibrium; that is, for whom the the new action satisfies the equilibrium conditions [I] and [II] of Section 3.6. Suppose moreover that $(a^*, 1 - a^*)$ is the expectation of $(x, 1 - x)$, conditional on $(x, 1 - x) \in P$. Then we say that there has been a *rational disruption* of the existing equilibrium convention, and moreover, that the types in P have *rationally deviated* from that convention.

3.8. **Rational Disruption: A Discussion.** Our notion of a rational disruption embodies the rational use of social expectations about the identity of "deviators." This is captured by two requirements: the deviators see their deviation as "equilibrium play" in the situation with a new action, and the public computes the expectation over all such deviator types to assign social credit.[15]

Rational disruption might be viewed as an equilibrium refinement. However, it differs importantly from standard refinements, which trim beliefs emanating from unplayed but universally available actions. We have here an action that is not available (initially at least) to most author pairs, but is, nevertheless, correctly interpreted by the academic public who assess credit. A rational disruption is then an equilibrium of a modified situation in which only a vanishingly small set of agents have access to the new action. This small set retains the full type distribution. They will deviate if and only if the public assessment of the new action makes it profitable to do so.[16]

The conceptual basis of the approach here is then different from that for standard refinements, but it may still be illuminating to contrast the two approaches. To do so, we proceed informally, supposing that there is a single player, Austen, in the first stage.[17] For simplicity, we restrict attention to the limiting case that $\delta = 0$, so that the signals are payoff-irrelevant.

In the Economics convention, where only α and ρ are used, Austen can readily be deterred from adopting the new action ⓡ by the accompanying belief that she made no contribution at all, so that $x = 0$. What implications would the "intuitive criterion" of Cho and Kreps (1987, p. 202) have here?[18] Does it rule out such "extreme" beliefs?

Consider then the set $S(ⓡ)$ of Austen's types who could not possibly gain from using ⓡ relative to the current equilibrium. However, *any* type $x \in [0, 1]$ could possibly gain from deviation, if

[15]Our concept is therefore related to neologism-proofness (Farrell 1993), in that it permits certain types of author pairs to profitably choose a fresh action, where the social evaluation of that action derives from the set of types who rationally deviate to that action.

[16]Hence a rational disruption is related to the notion of stability from evolutionary game theory. The definition of a rational disruption assumes that the size of the mutant group is *vanishingly* small. This simplifies the argument, but it could be replaced by the requirement that the size of mutant group is positive but *sufficiently* small, at the cost of greater complexity.

[17]After all, in the Economics convention with alphabetical default, it is Austen's preferences that are pivotal in determining the name order chosen in equilibrium. A more general analysis, one that allows for other conventions as well, would have to account for Byron's presence.

[18]The intuitive criterion is one of the best known equilibrium refinements. Cho and Kreps also provide a brief but clear survey of the entire literature.

the social assignment of credit following ⓡ is favorable enough.[19] Hence $S(ⓡ) = \varnothing$. Cho and Kreps then ask if any type *not* in $S(ⓡ)$ inevitably gains from using ⓡ, where the support of the beliefs involved excludes $S(ⓡ)$. However, *any* type can be made worse off after choosing ⓡ, if the beliefs this generates are unfavorable enough.[20] The intuitive criterion is therefore satisfied, and it has no effect in trimming relevant extreme beliefs here.

Our notion of "rational disruption" significantly restricts out-of-equilibrium beliefs relative to the intuitive criterion. Out-of-equilibrium beliefs that will dissuade Austen from deviating are not ruled out here by the intuitive criterion, as discussed. However, our requirement that there be a new equilibrium where ⓡ is available to a vanishingly small mutant group generates less unfavorable beliefs, derived from the set of types who actually choose ⓡ, and such a mutant group prospers.

4. The Economics Convention as an Equilibrium

In this section, we analyze the *Economics convention* E_t, which is alphabetical order, modified by name-reversal when $x < t$, for some threshold $t > 0$. Under this convention, using (3.1) applied to realized credits $(x, 1-x)$, Austen's utility from α is

$$A(\alpha, E) + \delta - \Gamma(z_B) = h(t) + \delta - \Gamma(h(t) - x), \tag{4.1}$$

while, using (3.2), her utility from ρ is

$$A(\rho, E) - \Gamma(z'_B) = l(t) - \Gamma(l(t) - x). \tag{4.2}$$

The only options available are alphabetical order α or reverse-alphabetical order ρ (and private randomizations over these schemes). We establish:

Theorem 1. *There is $t \in (0,1)$ such that the Economics convention E_t is an equilibrium.*

We relegate proofs to the appendix, but we include this particular proof in the text as it helps to understand the subsequent results.

[19] More precisely, Austen's payoff in the Economics convention is $l(\epsilon) - \Gamma(l(\epsilon)$ if $x \leq \epsilon$ and $h(\epsilon) - \Gamma(h(\epsilon) - x)$ if $x > \epsilon$. Either of these payoffs is less than or equal to $\max_{b\in[0,1]}[b - \Gamma(b-x)]$.

[20] That is, Austen's payoff in the Economics convention is $l(\epsilon) - \Gamma(l(\epsilon)$ if $x \leq \epsilon$ and $h(\epsilon) - \Gamma(h(\epsilon) - x)$ if $x > \epsilon$. Either payoff is greater than or equal to $\min_{b\in[0,1]}[b - \Gamma(b-x)]$.

Proof. We solve for the value of $x \in [0,1]$ at which Austen's preferences between α and ρ reverse, and then use a simple fixed point argument to ensure it coincides with society's anticipated threshold, t. From (4.1) and (4.2), observe that Austen will weakly prefer to reverse when

$$\Gamma\left(h(t)-x\right)-\Gamma\left(l(t)-x\right)\geq\Delta(t)+\delta, \tag{4.3}$$

where $\Delta(t)\equiv h(t)-l(t)$. By strict convexity of Γ, the left hand side is strictly decreasing in x, and so there exists a unique $x^*\geq 0$ such that Austen will strictly prefer to reverse if and only if x is smaller than x^*. In equilibrium, $x^*=t$, so using (4.3), t must solve

$$\Gamma\left(h(t)-t\right)-\Gamma\left(l(t)-t\right)=\Gamma\left(h(t)-t\right)=\Delta(t)+\delta, \tag{4.4}$$

whenever it is strictly positive. Noting that $m_A=h(0)$, Eq (3.3) guarantees such an t exists.

We claim that E_t is an equilibrium convention. When $x\in(t,1]$, Austen strictly prefers α to ρ (and all randomizations over α and ρ) because

$$\Gamma\left(h(t)-x\right)-\Gamma\left(l(t)-x\right)=\Gamma\left(h(t)-x\right)<\Gamma\left(h(t)-t\right)=\Delta(t)+\delta.$$

Therefore α is the equilibrium outcome, by [I] of Section 3.6.

When $x\in[0,t)$, Austen strictly prefers ρ to α because

$$\Gamma\left(h(t)-x\right)-\Gamma\left(l(t)-x\right)>\Gamma\left(h(t)-t\right)-\Gamma\left(l(t)-t\right)=\Gamma\left(h(t)-t\right)=\Delta(t)+\delta.$$

When $x\in[0,t]$, Byron also strictly prefers ρ to α, despite the possibility that he will now feel he is treating Austen unfairly. It is sufficient to consider the case in which Austen has contributed $x=t$ but only receives a credit of $l(t)$, in which case Byron receives the overall payoff $\delta+[1-l(t)]-\Gamma\left(t-l(t)\right)$ on reversal. Under α, he gets $1-h(t)$. Consequently, Byron will strictly prefer reversal at $x=t$ provided that

$$\Gamma\left(t-l(t)\right)<\Delta(t)+\delta. \tag{4.5}$$

Given (4.4) and Γ increasing, this condition holds if

$$h(t)-t>t-l(t). \tag{4.6}$$

So proving (4.6) completes the argument.[21] Suppose, on the contrary, that $h(t)-t \leq t-l(t)$. Then $t \geq [h(t)+l(t)]/2$ so $h(t)-t \leq [h(t)-l(t)]/2 = \Delta(t)/2$. Using (4.4), we must conclude that

$$\Gamma\left(\frac{\Delta(t)}{2}\right) \geq \Delta(t) + \delta. \tag{4.7}$$

Now, the function $\Omega(z) \equiv \Gamma(z/2) - z$ is convex, with $\Omega(0) = 0$. It follows that if $\Omega(z) > 0$ for some $z \in (0,1]$, then $\Omega(1) > 0$. It then follows from 4.7) that

$$\Gamma\left(\frac{1}{2}\right) > 1,$$

but this contradicts (3.4).

So, when $x \in [0,t)$, ρ is strictly preferred by both authors to α and therefore is the equilibrium outcome, by [II] of Section 3.6.[22]

□

The same result holds if private coordinated randomization is allowed. That is, suppose Austen and Byron agree to randomize name order by tossing a (possibly biased) coin at some $(x, 1-x)$. The expected utility (to Austen) of such a coin flip is sandwiched between the two utilities from α and ρ, so that if, say, the expected utility beats that from α, it must in turn be bettered by the utility from ρ. Generically (in x), such private randomization can never then occur.

Notice how the strict convexity of Γ and the conditions (3.3) and (3.4) are necessary to get what we see in practice. For instance, if Γ is linear, then Austen simply trades off units of her credit for Byron's. She will want to either always reverse, or never reverse. We see neither, which suggests that the "marginal loss" climbs as Byron's contribution climbs, for a fixed name order.

However, as already noted, the mere fact of being an equilibrium convention does not guarantee robustness to rational disruptions. We now turn to an examination of this question.

[21] The argument that follows is needed repeatedly in the Appendix, where it is given as Lemma 3.

[22] If $x = t$, Austen is indifferent between α and ρ, while Byron prefers reversal. This zero-probability case can be resolved either way.

5. DISRUPTING E_t WITH CERTIFIED RANDOM ORDER ⓡ

Certified random order ⓡ is an option that is institutionally provided, say by a consortium of the leading journals, so that the meaning of ⓡ is commonly known. The question is whether such certification can disrupt the Economics convention that utilizes α as the default but also involves the occasional reversal to ρ.

Given that the Economics convention is in effect, credits from the choice of α or ρ will continue to be given by (3.1) and (3.2), where t is pinned down by (4.4).[23] Should Austen and Byron employ random order for some realizations of x? If they do, a new pair of payoffs will be generated. These will consist of $\delta/2$ to each (in expected value), plus possibly asymmetric socially assigned credits, and any relevant loss terms. Of these three components, assigned credits will depend on society's view of just when the authors are agreeing to randomize. If, for instance, it is believed that they are doing so on an interval of x-realizations that is symmetric around $1/2$, and the density f is symmetric, then the credit will be split equally. The assignment of credit to previously unused strategies is restricted as in the notion of a rational disruption, as described in Section 3.7.

Theorem 2. *The equilibrium convention E_t is rationally disrupted by certified random order, ⓡ, once this option is introduced. Almost all the author pairs in this rational disruption who have ⓡ available and actually choose it are thereby made strictly better off.*

The Appendix provides a complete proof of this central result. This proof is involved, but we outline it here. Fix the equilibrium convention E_t, where t is given by (4.4). Suppose that society assigns an arbitrary social credit pair $(a, 1-a)$ to random order. For each such assignment, we find two thresholds defined on the domain of Austen's actual credit. One, which we call $x^\alpha(a)$, is such that Austen strictly prefers random order to alphabetical order if and only if her realized credit falls below $x^\alpha(a)$; (see Lemma 1). Another, called $x^\rho(a)$, is such that Austen strictly prefers random order to reverse order if and only if her realized credit lies above $x^\rho(a)$ (see Lemma 2). Over a subdomain of the a's, the former threshold lies above the latter, so there is a zone in which Austen strictly prefers random order to both alphabetical and reverse order. Moreover, by an intermediate value argument, there is a particular assignment of credit $a = a^*$ for which the conditional expected value of Austen's credit over this zone *equals* a^*.

[23] Recall that any rational disruption is adopted, at first, by a "small" fraction of the population of author pairs, so the payoffs ascribed to the author listings α and ρ are unaffected by the presence of the mutants.

Define $x_1^* \equiv x^\rho(a^*)$ and $x_2^* \equiv x^\alpha(a^*)$. The proof is completed by showing that in the zone (x_1^*, x_2^*), ⓡ is a rational deviation, as in Section 3.6, while it cannot be a rational deviation outside the zone $[x_1^*, x_2^*]$.[24] Recall that ⓡ is Austen's favorite outcome in (x_1^*, x_2^*). Moreover, we will also show that Byron strictly prefers random order in this range to alphabetical order (see Lemma 4). Therefore ⓡ strictly dominates the default and cannot be strictly dominated itself. Hence it is a rational deviation in (x_1^*, x_2^*), in the light of [I] and [II] of Section 3.6. In contrast, for $x > x_2^*$, Austen strictly prefers the default to ⓡ, so ⓡ cannot be a rational deviation, by [I] of Section 3.6. And for $x < x_1^*$, Austen strictly prefers ρ to ⓡ, and it can be shown that Byron does too (see Lemma 5). So ρ strictly Pareto dominates ⓡ, so that the latter cannot be an "equilibrium choice," by [II] of Section 3.6. We therefore have a rational disruption of the convention E_t.

6. Equilibrium for the Random Order Convention

We now analyze the *certified random order convention*, in which the action ⓡ is the default choice for either author. The set of available actions is now $\{d, \alpha, \rho\}$, where the default is $d =$ⓡ. The new option entails the two players randomizing with equal probability over the name orders α and ρ, with the ⓡ symbol attached to each realized outcome.[25]

Formally, a *certified random order convention* is described by two thresholds t and μ, such that $0 \leq t < \mu \leq 1$. If $x \in [t, \mu]$, the order of names is randomized, with the ⓡ symbol invoked for certification. The assumptions we have made will ensure that the randomization zone $[t, \mu]$ is nontrivial. The other two zones $[0, t)$ and $(\mu, 1]$ may or may not be nonempty. These are the "exception zones." In the first of these, Austen's contribution is small, and ρ is used. In the second, Austen's contribution is large, and α is used. Below, we show that at least one of these exception zones is nonempty.

If the distribution of contributions is symmetric, it will turn out that there is a symmetric equilibrium convention; that is, there is $t \in (0, 1/2)$ such that (i) if $x < t$ then the outcome is ρ, (ii) if $x > 1 - t \equiv \mu$ then the outcome is α, and (iii) if $x \in [t, \mu]$ then ⓡ is the outcome. Because at least one exception zone is nonempty, both exception zones are now nonempty, by symmetry.

[24] The end-points x_1^* and x_2^* have zero probability.

[25] Private correlated randomizations are also available, but they will never be used, and so we ignore them.

But there could be publicly observed situations — adviser-advisee pairs, research assistants, or the presence of a particularly reputable scholar — in which that symmetry is not to be had. In such situations we impose the following additional restriction. Recall that Γ is strictly increasing and continuously differentiable. Define:

$$G = \inf_{z>0} \frac{\Gamma'(z)z}{\Gamma(z)}.$$

Because Γ is strictly convex, $G \geq 1$. (For instance, if $\Gamma(z) = z^k$ for $z > 0$ and some $k > 1$, then $G = k$.) We assume that the density function of contributions f is such that the following condition is satisfied: for every pair (t, t') with $t < t'$,

$$l(t') - l(t) \leq G[t' - e] \text{ and } h(t') - h(t) \leq G[t' - t]. \tag{6.1}$$

It is not hard to see that there exists a non-empty set of f for which these conditions hold.[26]

6.1. **Equilibrium With Certified Random Order.** We maintain the description of co-author interaction from Section 3.6. Specifically, contributions $(x, 1-x)$ are first revealed. Next, Austen and Byron interact. It is presumed that no outcome to which either player strictly prefers the default can be an equilibrium. Further, if any outcome is strictly Pareto-dominated, it cannot be an equilibrium outcome.

Theorem 3. *If f satisfies* (6.1)*, there exists an equilibrium random-order convention with thresholds (t, μ), where $t < \mu$, so that ⓡ is always used over a range of relative contributions. Moreover,* either $t > 0$ or $\mu < 1$ or both*, so that at least one of the exception zones is nonempty.*

If f is symmetric, there exists a symmetric equilibrium random-order convention where $0 < t < \mu = 1 - t < 1$, so randomization, alphabetical order and reverse alphabetical order are all used under the convention.

[26]The simplest example is that of a uniform density: $f(x) = 1$ for all $x \in [0, 1]$. In this case, $l(t') - l(t) = (1/2)(t' - t) < G(t' - t)$, since $G \geq 1$. Similarly, $h(t') - h(t) = (1/2)(t' - t) < G(t' - t)$. This condition suggests that density functions that are sufficiently close to uniform will also work. More precisely, consider (6.1) as it applies to l; the argument for h is analogous. Since l is differentiable, it is sufficient to show that $\frac{dl}{dt} \leq G$. We have $l(t) = \frac{\int_0^t xf(x)dx}{F(t)}$ so that $\frac{dl}{dt} = \frac{f(t)}{F(t)^2}\left[tF(t) - \int_0^t xf(x)dx\right]$. Suppose that the density function is bounded above and below so that $f(x) \in [\underline{f}, \overline{f}]$, for all $x \in [0, 1]$, where $\overline{f} \geq 1 \geq \underline{f} > 0$. Since $F(t) \leq \overline{f}t$ and $\int_0^t xf(x)dx \geq \underline{f}t^2/2$, it follows that $\frac{dl}{dt} \leq y^2 - y/2$, where $y = \frac{\overline{f}}{\underline{f}}$. Hence $\frac{dl}{dt} \leq G$ if $\frac{\overline{f}}{\underline{f}} \leq 1/4 + \sqrt{G + 1/16}$. That is, Eq (6.1) is satisfied for all distributions whose density functions are suitably bounded above and below. The range of bounds is always non-empty, and becomes larger, with larger G.

The Appendix contains a detailed proof of existence; here is an outline of the argument. Suppose that certified randomization carries the credits $(m, 1-m)$, where m can be shown to be the conditional expectation of x over a non-empty interval that contains m. For all contributions, x, by Austen that are smaller than m, define a function t where $t(m)$ is Austen's indifference threshold for reverse order ρ (set $t(m) = 0$ if she never wishes to switch). Likewise, for all contributions by Austen larger than m, let $\mu(m)$ be the analogous threshold for Byron for switching to alphabetical order α. Our assumptions on f guarantee that t and μ are uniquely defined and continuous in m. Therefore the mapping

$$m \mapsto m' \equiv \mathbb{E}(x|x \in [t(m), \mu(m)])$$

is well-defined and continuous and so admits a fixed point m^*. Let $t^* = t(m^*)$ and $\mu^* = \mu(m^*)$.

To prove that such a convention is an equilibrium, consider first the exception zone $[0, t^*)$, provided it is nonempty. By construction of our fixed point, Austen strictly prefers ρ to ⓡ in this region. But we show this is true of Byron as well. Indeed, ρ is Byron's favorite outcome when $x \leq t^*$. Thus ρ strictly Pareto-dominates the default and cannot be strictly Pareto-dominated itself, so it is an equilibrium outcome in this range, by [I] and [II] of Section 3.6. By an analogous argument, α is an equilibrium outcome when x lies in the exception zone $(\mu^*, 1]$.

To complete the argument, consider the randomization zone $[t^*, \mu^*]$. In the sub-region $(t^*, m^*]$, the default ⓡ is Austen's favorite outcome, so it is the equilibrium outcome, by [I] of Section 3.6. (At t^*, it remains a possible equilibrium outcome.) A similar argument involving Byron holds in the sub-region $[m^*, \mu^*]$.

6.2. **Do Rational Disruptions Exist?** Consider now the possibility of rational disruptions from the equilibria established in Theorem 3. By the definition of equilibrium, it is clear there can be no rational disruption involving a name-order that is already in use. What if an exception zone is empty? Suppose, for example, that $0 = t < \mu < 1$, so that the name order ρ is not used for Austen and Byron. Could there be a rational disruption based on ρ? The following theorem describes the possibilities:

Theorem 4. *There can be no rational disruption of the equilibrium random-order convention involving a name scheme already in use. Furthermore, if $t = 0$, there can be no rational disruption*

to ρ that involves any $x < m^*$; similarly if $\mu = 1$, there can be no rational disruption to α that uses any $x > m^*$.

Theorem 4 states that ⓡ is robust to rational disruptions that retain the "natural meaning" of the name order used in the disruption. First, if both exception zones are nonempty — this is true in the symmetric case — stability of the equilibrium to all disruptions is guaranteed, since all name orders have established equilibrium meanings.

Now suppose that one of the exception zones — say the one that involves ρ — is empty. That is, the order Byron-Austen is never observed, owing to some asymmetry in f. However, it is reasonable to suppose there are other author pairs with a similar asymmetry but in the reverse order. Suppose Charlotte Bronte and W.H. Auden are such a reversed pair. That is, Bronte's $1 - x$ in the Bronte-Auden pair is distributed the same way as Austen's x in the Austen-Byron pair. Since the order Austen-Byron *is* observed, so too must the order Bronte-Auden be observed, since the random-order convention here has no intrinsic alphabetical bias. The Bronte-Auden order implies that Bronte contributed the lion's share. That is, a reversal of the alphabetic order has the "natural meaning" that the author who is now first did most of the work. Hence, although the name order ρ does not arise for Austen and Byron, it arises elsewhere (for Bronte and Auden) and so it has an established meaning. We restrict the meaning of the unused ρ for Austen and Byron to be that x (Austen's contribution) is small relative to $1 - x$.

Indeed, Theorem 4 states that no rational disruption by ρ can involve any $x < m^*$, where m^* is the mean contribution for Austen, conditional on being in the randomization zone. Hence Austen's contribution cannot be small relative to that of Byron when such a surprise deviation to ρ is observed.[27]

6.3. **Efficiency Gain From the Random Order Convention: An Example.** So far, we have shown that the Economics convention is an equilibrium with the action set $\{\alpha, \rho\}$ but is subject to rational disruption using the name order ⓡ, once this option is introduced. On the other hand, the random order convention is an equilibrium and is not subject to rational disruption from the

[27]The reason that there is no rational disruption by ρ that preserves the natural meaning of ρ is that the original equilibrium was constructed allowing a role for ρ to signal a disproportionate contribution by Byron, but this role was not needed. Interestingly, we cannot rule out a rational deviation to ρ that gives ρ a completely new interpretation—signaling that the contributions are intermediate between those for ⓡ and those for α.

set $\{ⓡ, \alpha, \rho\}$, if these orders are already used. If α or ρ is not used, there cannot be a rational disruption that respects the natural meanings of these orders. Put another way, no intervention is necessary to break out of the equilibrium E_t convention or to remain in the ⓡ convention. In this section, we show further that the equilibrium ⓡ convention may generate higher aggregate welfare and is not simply fairer.

There is a clear basis for an efficiency advantage of ⓡ over the Economics convention. The social credit and the pure gain from first authorship, δ, are both constant-sum components of the two players' payoffs. The loss function terms, however, are not constant-sum. Certified random order may then reduce loss on average since it uses three signals instead of two, permitting signals for exceptional contributions for both Austen and Byron. That could then reduce the extreme values of loss, and hence the average values as well. But this intuition is incomplete: it is *a priori* possible for the three ranges under ⓡ to be so badly situated that the total expected payoff under E_t is greater. We therefore explore the issue further by considering, as an example, the analytically tractable case of a uniform density of contributions.

Theorem 5. *Assume that f is uniform on $[0, 1]$. Then a symmetric random order convention using ⓡ is more efficient than the Economics convention E_t, in the sense of having a strictly higher sum of expected overall payoffs for the two agents.*

In this special case, at least, the case for ⓡ does not rely on considerations of fairness. The sum of the two agents' expected utilities is higher under ⓡ than under E, so that *any* symmetric quasiconcave welfare criterion (including Bentham's additive utilitarianism) would strictly prefer ⓡ to E.[28]

7. Remarks and Extensions

7.1. **Private Versus Certified Randomization.** The mechanism of *private* randomization is not a novel one. In the simplest case, this involves two researchers flipping a coin to decide the order of names and — in its most effective form — including a lead footnote to that effect. This has been used previously on a number of occasions. The present mechanism of certified randomization

[28]We assume that such a welfare criterion is defined on the *expected* utilities of the agents. This is appropriate if the "publication game" is repeated often, so that there are many independent draws of x.

differs from such private randomization. In particular, the exact comparison of the two mechanisms depends on the channel through which published papers come to the attention of other researchers.

The present paper is motivated by the key channel of *written* citations to the paper in the subsequent literature. In this case, certified randomization outperforms private randomization. What is important here is that the symbol ⓡ is maintained in subsequent references, whereas the lead footnote with private randomization is not. This implies that private randomization can generate only two permanently observed configurations for Austen and Byron—α or ρ, which limits its effectiveness.

But there are other channels. The most obvious is direct viewing of the paper by another researcher. Given that there is a lead footnote indicating private randomization, presumably noted by the researcher, as is the symbol ⓡ under our mechanism, the two mechanisms are essentially equivalent, if direct viewing is the only channel. That is, there are four possible outcomes of a collaboration between Austen and Byron under private randomization: α, ρ, $\alpha^{\text{Ⓕ}}$ or $\rho^{\text{Ⓕ}}$, where $\rho^{\text{Ⓕ}}$, for example, means that the authors are listed as first Byron, then Austen, and there is a lead footnote. These possibilities correspond precisely to the four possibilities under the ⓡ mechanism, with only notational differences. But the correspondence is weakened as written citations to the paper co-exist with direct readings.

The last channel that seems worthy of mention concerns verbal and written allusions to the published paper in seminars. What happens here depends on how exactly this citation is made. If the symbol ⓡ is retained (e.g., on slides), whereas the lead footnote escapes attention, the advantages of our mechanism over private randomization remain. If no mention is made of ⓡ, the effectiveness of our mechanism — at least along this one dimension — will be reduced to match that of private randomization. Certified randomization is always, then, at least as effective as private randomization, and, for at least one important channel, that of written citations in subsequent literature, strictly more effective.

7.2. **Partial Adoption by Journals.** Our model shows that a small group of coauthors can successfully invade the E_t convention. What if adoption by journals were also incomplete to begin with? What if the *American Economic Association*, for instance, threw its weight behind this new

scheme, but other journals did not? If articles that were published in the *American Economic Review* with the ⓡ symbol were still referenced in other journals complete with the new symbol, our analysis would apply with minimal reinterpretation. That is, partial adoption in these circumstances by journals would simply serve to scale down the effective size of the group using the new convention, without changing the payoffs.

If other journals declined to print the symbol ⓡ in their references, payoffs to an invading group would be modified. These payoffs would now reflect a combination of the payoffs from a correct interpretation of the symbol ⓡ, and the old Economics equilibrium convention, since ⓡ is subsequently lost. If Austen knows that *nobody* will reproduce the ⓡ in their citations, use of this option is tantamount to her randomizing across α and ρ. Such randomization could occur at the reversal threshold t, and nowhere else. If more generally, a small fraction of citations respects the new symbol in citations, then the fixed point argument used to obtain a rational disruption zone should work much as it did before. This could generate a smaller disruption zone, but one that will still overlap the threshold t.

It would be useful if the AEA not merely adopted this new convention, but also used their influence to pressure other journals adopt it as well, or at least to respect the new style by including ⓡ in references to AEA papers. However, the effect of the AEA using its influence like this would merely be to speed up the evolution towards the new system; wider acceptance does not seem crucial to ultimate success.

7.3. **Other Conventions And Actions.** In a world where contributions $(x, 1-x)$ are common knowledge to the two authors, there are alternative conventions that can achieve higher degrees of efficiency, and mutants built along those lines could invade the random order convention.

Indeed, there is a formal mechanism that attains full efficiency, as follows. Suppose that for each $x < 1/2$, ρ is used and ⓧ is appended to the names, whereas, if $x \geq 1/2$, α is used and ⓧ is appended. Neither agent then ever experiences any loss, so that overall expected payoffs are $1+\delta$, which is the upper bound. Moreover, there can be no disruption of this convention that both authors would participate in.

But such a mechanism pushes very hard the assumption that the agents have common knowledge of x. Presumably, agreement would be elusive and there would be endless bitter arguments about

the exact value of x. Consider, on the other hand, a solitary pair of authors who disagree about the value of x in the context of the random order convention. In the first place, even if these authors disagree about the exact value of x, it is enough that they agree it is in the range where a particular name order is chosen. It is, furthermore, helpful that, in the random order convention, the default can only be overturned by mutual consent. That is, even if Austen, for example, believes that x warrants the naming scheme α whereas Byron believes that x warrants the scheme ⓡ, it is at least clear to both authors that ⓡ will be chosen.

7.4. **Randomizing Citations.** An alternative to our proposed mechanism would be to keep published papers with the authors names' listed alphabetically, but to randomize 50-50 each time a *citation* is made.[29] Such a randomization might be strictly socially optimal even with social indifference between the two name orders. This would reflect randomization being assessed as a fair means for allocating an indivisible item, as with "Machina's Mom" (Machina 1989). We believe that the ⓡ mechanism has advantages over this scheme.

In the first place, as a practical matter, it would be hard to ensure that all researchers citing the paper diligently randomize, as might be especially true in seminar presentations. Although it might be possible to cite the famous paper as Douglas-Cobb instead of Cobb-Douglas, for example, it seems it would be difficult to cite it sometimes as Cobb-Douglas and sometimes as Douglas-Cobb.[30]

Moreover, this alternative mechanism does not allow co-authors to indicate the infrequent (but by no means exceptional) situation in which one of them has done the lion's share of the work. For example, the order of names describing the Stolper-Samuelson theorem signalled the greater contributions of Stolper, as graciously acknowledged by Samuelson. This possibility would be lost under the randomization of citations but is preserved — and made symmetric — under the certified random order convention.

7.5. **Strategic Authorship Decisions.** Authors may choose whom to co-author with, given the going convention. For instance, later authors may be more reluctant to engage in projects with multiple co-authors, for fear of falling into *et al* oblivion. One might also make the converse argument: that under alphabetical order, early authors are more willing to offer co-authorship to

[29]Leeat Yariv proposed this device, perhaps as a supplement to the mechanism here.

[30]However, it would be useful to find some way of equalizing credit for *past* publications, where ⓡ cannot be retroactively imposed.

late authors, knowing that this will have only a small effect on their payoffs, being listed first anyway.[31] Indeed, Austen might be excessively eager to offer co-authorship to Zeno, anticipating that "Austen and Byron" would now be transformed into "Austen *et al*." Einav and Yariv (2006) find some evidence for both these effects: late authors are more likely to be involved in publications that involve non-alphabetical orderings; the effect is particularly strong when there are three or more authors. It is also the case that early authors are more likely to be involved in four- or five-author projects. A full accounting of these and other strategic factors in the selection of co-authors demands a model; one of us has indeed written down a set of notes to this effect; Ray (2013).

7.6. **Three Or More Authors.** The entire analysis in this paper has been for the case of two authors. While we foresee no great difficulty in extending the analysis to the case of three or more authors, there are additional complications that will need to be addressed. For instance, one possible choice would be partial randomization of the form:

[Zeno ⓡ Byron] and Austen.

The best initial approach might be to rule out such possibilities and restrict attention to randomization over the entire list; e.g.,

Zeno ⓡ Austen ⓡ Byron.

The extension of the analysis in this paper to such conventions should then be relatively straightforward.

8. Conclusions

In this paper we describe a scheme — certified random order — for assigning credit to papers with two coauthors. We first characterize the current system of joint authorship as modified-alphabetical, where the author who is earlier in the alphabet can offer first authorship to the other, if the contributions are very unequal. This is motivated by a loss term for the earlier author.

The new scheme involves flipping a coin to determine first authorship and adding the notation ⓡ to the list of the two authors when this has been done. In addition, we allow either author to

[31] We are grateful to Sahar Parsa and Phil Reny, both lexicographically challenged and clearly on the lookout for such dangers, for this point.

offer first authorship to the other, without the ⓡ notation, again motivated by a loss term when the contributions have been extremely unequal.

We show that if such a scheme is made available, then it will enter into existing society via a "rational disruption" of the existing convention based on alphabetical order. On the other hand, there is no possibility of reverting to alphabetical order once in the new convention comes to dominate. In short, we do not seek to impose such a system. We only claim that if it is offered, it will be adopted. Moreover, we show that the new equilibrium convention may generate a higher sum of expected utilities than does the old. The new mechanism would then be strictly preferred on the basis of efficiency criteria, in addition to the core principle of fairness across authors.

The beauty of the mechanism ⓡ is that it does not demand any more of the agents than does the present Economics convention E. The convention ⓡ simply allows *either* player to concede first authorship, instead of allowing only the first author to have this option, as in the convention E. Although such an option can lead to arguments, it is indeed exercised on occasion in reality.[32]

The analysis in this paper focuses on equilibrium conventions, ones where the alphabetical order prevails or where certified random order prevails. A more complete analysis would examine the full dynamical system in which both systems could co-exist. In the transition from an old to a new convention, the default choice would have to switch at some point from the old convention default to the new. The key issue is to model how this transition might occur.

In summary, certified random order: (a) distributes the gain from first authorship evenly over the alphabet, (b) allows *either* author to signal credit when contributions are extremely unequal, (c) will be willingly adopted even in an environment where alphabetical order is the default, (d) is robust to deviations, (e) may dominate alphabetical order on the grounds of ex-ante efficiency, and (f) with the minor exception of a simple symbol, it is no more complex than the old system.

REFERENCES

Carney, D. R. and Banaji, M. R. (2012), "First is Best," *PLoS ONE* **7**(6), e35088. doi:10.1371/journal.pone.0035088

[32]Flipping a coin and adding ⓡ to the list of authors also entail costs, but these seem trivial.

Chambers, R., Boath, E., and Chambers, S. (2001), "The A to Z of Authorship: Analysis of Influence of Initial Letter of Surname on Order of Authorship," *British Medical Journal* 323(22-29 Dec.), 1460–1461. doi:10.1136/bmj.323.7327.1460

Cho, I.-K. and D. Kreps (1987), "Signaling Games and Stable Equilibria" *Quarterly Journal of Economics* **102**, 179-221.

Einav, L. and L. Yariv (2006), "What's in a Surname? The Effects of Surname Initials on Academic Success," *Journal of Economic Perspectives* **20**, 175–188.

Engemann, K. and H. Wall (2009), "A Journal Ranking for the Ambitious Economist," *Federal Reserve Bank of St. Louis Review* **91**, 127–39.

Engers, Maxim; Gans, Joshua S.; Grant, Simon; and King, Stephen P. (1999) "First-Author Conditions," *Journal of Political Economy* 107, 859-883.

Farrell, Joseph. (1993). "Meaning and Credibility in Cheap-Talk Games," *Games and Economic Behavior* **5**, 514-531.

Feenberg, Daniel R., Ganguli, Ina, Gaule, Patrick and Jonathan Gruber (2015) "It's Good to be First: Order Bias in Reading and Citing NBER Working Papers," NBER Working Paper No. 21141.

Haque, A., and Ginsparg, P. (2009), "Positional Effects on Citation and Readership in arXiv," *Journal of the American Society for Information Science and Technology* **60** (11), 2203–2218. doi:10.1002/asi.21166

Itzkowitz, J., Itzkowitz, J., and Rothbort, S. (2016), "ABCs of Trading: Behavioral Biases Affect Stock Turnover and Value," *Review of Finance* **20**, 663–692.

Jacobs, H. and Hillert, A. (2016), "Alphabetic Bias, Investor Recognition, and Trading Behavior," *Review of Finance* **20**, 693–723.

Machina, M. (1989), "Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty," *Journal of Economic Literature* **27**, 1622-1668.

Ray, D. (2013), "All the Names: Some Strategic Consequences of Alphabetical Order in Joint Research," mimeo., New York University.

van Praag, C. M. and van Praag, B. M. S. (2008), "The Benefits of Being Economics Professor A (Rather than Z)," *Economica* **75**, 782–796. doi:10.1111/j.1468-0335.2007.00653.

Weber, Matthias (2016) "The Effects of Listing Authors in Alphabetical Order: A Survey of the Empirical Evidence."
SSRN: http://ssrn.com/abstract=2803164, or http://dx.doi.org/10.2139/ssrn.2803164

9. APPENDIX: PROOFS

Proof of Theorem 2. Fix an equilibrium convention E_t with its associated reversal threshold $t > 0$, given by (4.4), and suppose that society assigns a credit pair $(a, 1-a)$, for $a \in [0,1]$, to the observation of random order. We begin with a lemma that compares random order to α for Austen.

Lemma 1. *There exists $\bar{a} \in (l(t), h(t))$ with the following properties: there is a continuous function x^α such that, for all $a \in [l(t), \bar{a}]$, $x^\alpha(a) \in [0,1]$, and Austen strictly prefers random order over α whenever realized contributions $(x, 1-x)$ satisfy $x \in [0, x^\alpha(a))$, and strictly prefers α to random order when $x \in (x^\alpha(a), 1]$. Moreover,*

$$x^\alpha(l(t)) > t > l(t) > 0, \tag{9.1}$$

and

$$x^\alpha(a) > a \text{ for all } a \in [l(t), \bar{a}) \text{ with } x^\alpha(\bar{a}) = \bar{a}. \tag{9.2}$$

Proof. Random order at realization $(x, 1-x)$ yields an expected payoff to Austen of

$$a + \frac{\delta}{2} - \Gamma(a - x) \tag{9.3}$$

while alphabetical order generates a payoff of

$$h(t) + \delta - \Gamma(h(t) - x)$$

as described in (4.1). Therefore random order is weakly preferred to α if

$$\Gamma(h(t) - x) - \Gamma(a - x) + a \geq h(t) + \frac{\delta}{2}. \tag{9.4}$$

If this inequality holds for some x, then *equality* must hold for some $x^\alpha(a)$, because for x large enough the inequality (9.4) is strictly reversed.[33] If (9.4) fails for all x, we formally set $x^\alpha(a) = 0$. Because $\Gamma(z)$ is strictly convex when $z > 0$, the LHS of (9.4) is *strictly* decreasing in x. Supposing, for the moment, that $x^\alpha(a) > 0$, this shows that Austen will strictly prefer random order to α

[33]Recall that $\Gamma(z) = 0$ for all $z \leq 0$, and is continuous everywhere. This, combined with $a \leq h(t)$, guarantees that (9.4) must fail for x large enough.

when $x \in [0, x^\alpha(a))$, will be indifferent at $x^\alpha(a)$, and will strictly prefer α to random order when $x \in (x^\alpha(a), 1]$.

To establish (9.1), set $a = l(t)$ and $x = t$. Then

$$\begin{aligned}\Gamma\left(h(t) - x\right) - \Gamma\left(a - x\right) + a &= \Gamma\left(h(t) - t\right) - \Gamma\left(l(t) - t\right) + l(t) \\ &= \Gamma\left(h(t) - t\right) + l(t) \\ &= \Delta(t) + \delta + l(t) > h(t) + \frac{\delta}{2},\end{aligned}$$

where the third equality employs the definition of t in (4.4).[34] That shows that (9.4) holds as a strict inequality when $x = t$. Since the left-hand side of (9.4) is decreasing in x, (9.1) is true.

It follows that x^α is continuous in a, whenever $x^\alpha(a) > 0$. Indeed, it is continuous always given the formal assumption that $x^\alpha(a) = 0$ if (9.4) fails for all x. Since (9.4) must fail for every x when $a = h(t)$, it follows that $x^\alpha(h(t)) = 0$. By the Intermediate Value Theorem, there exists $\hat{a} \in (l(t), h(t))$ such that $x^\alpha(\hat{a}) = \hat{a}$. Because $x^\alpha(a)$ is continuous on $[l(t), h(t)]$ and $x^\alpha(l(t)) > l(t)$, there is a *smallest* such $\hat{a}$; call it $\bar{a}$. It must be that $x^\alpha(a) > a$ for all $a \in [l(t), \bar{a})$, which establishes (9.2) and completes the proof. □

Our next lemma establishes a corresponding threshold for the comparison of random order and reverse-alphabetical order ρ. We will work on the domain $[l(t), \bar{a}]$.

Lemma 2. *There is a continuous function $x^\rho : [l(t), \bar{a}] \to [0, 1]$ such that Austen strictly prefers random order to ρ if $x \in (x^\rho(a), 1]$, and strictly prefers ρ to random order if $x \in [0, x^\rho(a))$. Moreover,*

$$x^\rho(l(t)) = 0, \text{ and } x^\rho(a) < a \text{ for all } a \in [l(t), \bar{a}]. \tag{9.5}$$

Proof. Reverse order ρ yields a payoff to Austen given by (4.2), which is

$$l(t) - \Gamma\left(l(t) - x\right).$$

[34] The credits are assigned are the same as in ρ, and Austen is indifferent between α and ρ at t. So she will strictly prefer random order, which yields the same credit to Byron but gives Austen the extra payoff of δ half the time.

Combining with (9.3), we see that random order is weakly preferred to ρ if

$$\Gamma(a-x)-\Gamma(l(t)-x)-a\leq\frac{\delta}{2}-l(t). \tag{9.6}$$

Again, because $\Gamma(z)$ is strictly convex when $z>0$, the LHS of (9.6) is strictly decreasing in x, thereby showing that (9.6) holds over some interval of the form $[x^\rho(a),1]$. Note that (9.6) must hold, in particular, at $x=1$, since $a\geq l(t)$. Hence $x^\rho(a)$ is either zero, if (9.6) always holds, or it is the value $x^\rho(a)>0$ for which (9.6) holds with equality.

Finally, we establish (9.5). When $a=l(t)$, the same social credits are associated with random order as with reversal. So the value of Γ at $a=l(t)$ is the same whether Austen reverses or randomizes, but in the latter case she also picks up the δ payoff with probability $1/2$. So Austen will strictly prefer random order. That is, at $a=l(t)$, (9.6) holds for all $(x,1-x)$, so $x^\rho(l(t))=0$.

To establish the second part of (9.5), suppose that at $(a,1-a)$, the realized contributions are *also* $(a,1-a)$. Setting $x=a$, we see that

$$\Gamma(a-x)-\Gamma(l(t)-x)-a=\Gamma(0)-\Gamma(l(t)-a)-a=-a\leq-l(t)<\frac{\delta}{2}-l(t),$$

which shows that (9.6) is satisfied with strict inequality when $x=a$.[35] Therefore $x^\rho(a)<a$. □

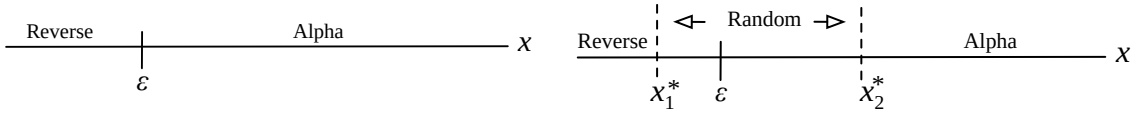
We now use the previous two lemmas to derive an equilibrium value of socially assigned credit, a^*. Let $\phi(a)$ denote the conditional expected contribution by Austen over all values of x for which Austen prefers random order to α and ρ. This is the expectation of x conditional on x lying in the interval $[x^\rho(a),x^\alpha(a)]$. Lemmas 1 and 2 together tell us that $x^\rho(a)<a\leq x^\alpha(a)$ whenever $a\in[l(t),\bar{a}]$, so ϕ is defined on $[l(t),\bar{a}]$. Because x^α and x^ρ are continuous, so is ϕ. We know from (9.1) that $x^\alpha(l(t))>t$, and we know from (9.5) that $x^\rho(l(t))=0$, so it follows that

$$\phi(l(t))=\mathbb{E}(x|x\leq x^\alpha(l(t)))>\mathbb{E}(x|x\leq t)=l(t). \tag{9.7}$$

We also know from (9.2) that $x^\alpha(\bar{a})=\bar{a}$, so that

$$\phi(\bar{a})\leq\bar{a}. \tag{9.8}$$

[35]The weak inequality in the chain uses the fact that $a\geq l(t)$ and the assumption that $\Gamma(z)=0$ when $z\leq 0$.



(a) Existing Regime **(b)** Deviation to Random Order

Figure 1. Incentives to Deviate to Random Order

Combining (9.7) and (9.8) and invoking the continuity of ϕ, it follows there exists $a^* \in (l(t), \bar{a}]$ such that

$$\phi(a^*) = a^*.$$

Define $x_1^* = x^\rho(a^*)$ and $x_2^* = x^\alpha(a^*)$. See Figure 1. Austen prefers random order ⓡ to either α or ρ whenever x lies in the interval $[x_1^*, x_2^*]$ (strictly so in the interior), provided the public assigns credit of $(a^*, 1 - a^*)$.

The following minor technical result, already used in the proof of Theorem 1, is needed repeatedly in the following proofs as well.

Lemma 3. *Suppose that* $\Gamma(z_1) > z_2$ *for some* $z_1 \in (0, 1]$ *and* $z_2 \in [0, 1]$. *Then* $z_2 - z_1 < z_1$.

Proof. Suppose, on the contrary, that $z_2 - z_1 \geq z_1$. Then $z_1 \leq z_2/2$, and because Γ is increasing,

$$\Gamma(z_2/2) \geq \Gamma(z_1) > z_2. \tag{9.9}$$

Now, $\Omega(z) \equiv \Gamma(z/2) - z$ is convex, and $\Omega(0) = 0$. Therefore, if $\Omega(z) > 0$ for some $z \in (0, 1]$, it must be the case that $\Omega(1) > 0$. But (9.9) implies that $\Omega(z_2) > 0$. Moreover, $z_2 \geq 2z_1 > 0$ and $z_2 \leq 1$. It follows that $\Gamma(1/2) > 1$, but that contradicts (3.4). □

We now apply Lemma 3 to prove the next two lemmas that establish key properties of Byron's preferences.

Lemma 4. *Byron strictly prefers* ⓡ *to* α *in the region* $[x_1^*, x_2^*]$.

Proof. Note that Byron receives at most the payoff $1 - h(t)$ under α, while under random order his lowest possible expected payoff is $(1 - a^*) + (\delta/2) - \Gamma(x_2^* - a^*)$ (this corresponds to the highest contribution x_2^* by Austen for which Austen prefers random order over α). Comparing these, it is

sufficient to show that

$$\frac{\delta}{2} + (h(t) - a^*) > \Gamma(x_2^* - a^*). \tag{9.10}$$

Recall that Austen herself is indifferent over the two options at x_2^*, so that

$$\delta + h(t) - \Gamma\left(h(t) - x_2^*\right) = \frac{\delta}{2} + a^* - \Gamma(a^* - x_2^*) = \frac{\delta}{2} + a^*,$$

(where the second equality comes from $x_2^* = x^\alpha(a^*) \geq a^*$ by (9.2)). Transposing terms,

$$\frac{\delta}{2} + [h(t) - a^*] = \Gamma\left(h(t) - x_2^*\right). \tag{9.11}$$

Define $z_1 \equiv h(t) - x_2^*$ and $z_2 \equiv h(t) - a^*$. Since $a^* \leq \bar{a} < h(t)$ and (9.11) holds, the conditions of Lemma 3 are met. It follows that $x_2^* - a^* < h(t) - x_2^*$. Combining this inequality with (9.11) and recalling that Γ is increasing, we obtain (9.10). □

Lemma 5. *Byron strictly prefers ρ to ⓡ when $x < x_1^*$.*

Proof. Under random order, Byron's payoff is $(1 - a^*) + (\delta/2)$, there being no additional loss because $x < x_1^* < a^*$. Under ρ, Byron's lowest payoff occurs when $x = x_1^*$, and it is given by $[1 - l(t)] + \delta - \Gamma(x_1^* - l(t))$. So it is sufficient to show that

$$a^* - l(t) + \frac{\delta}{2} > \Gamma(x_1^* - l(t)). \tag{9.12}$$

Because $x < x_1^*$, we have $x_1^* > 0$, so that Austen is indifferent between ρ and ⓡ at x_1^*. Therefore

$$a^* - l(t) + \frac{\delta}{2} = \Gamma(a^* - x_1^*) - \Gamma(l(t) - x_1^*).$$

Because $a^* > l(t)$, (9.12) is trivially true when $x_1^* \leq l(t)$. So we suppose that $x_1^* > l(t)$, in which case the above equality can be written as

$$a^* - l(t) + \frac{\delta}{2} = \Gamma(a^* - x_1^*). \tag{9.13}$$

Define $z_1 \equiv a^* - x_1^*$ and $z_2 \equiv a^* - l(t)$. Since $a^* \geq x_1^* \geq l(t)$ and (9.13) holds, the conditions of Lemma 3 are met. It follows that $a^* - x_1^* > x_1^* - l(t)$. Combining this inequality with (9.13) and recalling that Γ is increasing, we obtain (9.12). □

We can now complete the proof of the theorem. First we show that in the zone (x_1^*, x_2^*), ⓡ is an equilibrium outcome. Recall that ⓡ is Austen's favorite outcome in this range. By Lemma 4, Byron strictly prefers random order in this range to alphabetical order. Therefore ⓡ strictly Pareto-dominates the default and is not strictly Pareto-dominated itself, so it is a rational deviation for all types in (x_1^*, x_2^*), by [I] and [II] of Section 3.6. (The zero probability endpoint $x = x^*$ can be attached to either ⓡ or ρ, and a similar assertion holds for $x = x_2^*$.) Next, for $x > x_2^*$, Austen strictly prefers the default α to ⓡ, so ⓡ cannot be a rational deviation, by [I] of Section 3.6. For $x < x_1^*$, Austen strictly prefers ρ to ⓡ by the construction of x_1^*, and Byron does too, by Lemma 5. Therefore ρ strictly Pareto dominates ⓡ, which means that the latter cannot be a rational deviation, by [II] of Section 3.6. We have therefore established that E_t is rationally disrupted by random order. □

Proof of Theorem 3. We begin by setting up a particular random order convention (t^*, μ^*), and then show it is an equilibrium. To this end, define for any $m \in [0, 1]$ and $t \in [0, m]$,

$$\Psi_A(t, m) \equiv [m - l(t)] + \frac{\delta}{2} - \Gamma(m - t), \tag{9.14}$$

and for any $\mu \in [m, 1]$,

$$\Psi_B(\mu, m) \equiv [h(\mu) - m] + \frac{\delta}{2} - \Gamma(\mu - m). \tag{9.15}$$

First consider the general case in which f satisfies (6.1). The following lemma helps to obtain the equilibrium socially assigned credit, m^*.

Lemma 6. *Assume that* (6.1) *holds. Then whenever* $\Psi_A(t, m) \leq 0$, $\Psi_A(t, m) > \Psi_A(t', m)$ *for all* $t' < t$, *and whenever* $\Psi_B(\mu, m) \leq 0$, $\Psi_B(\mu, m) > \Psi_B(\mu', m)$ *for all* $\mu' > \mu$.

Proof. Suppose that $\Psi_A(t, m) \leq 0$. Pick $t' < t$. Then

$$\begin{aligned}
\Psi_A(t, m) - \Psi_A(t', m) &= [\Gamma(m - t') - \Gamma(m - t)] - [l(t) - l(t')] \\
&\geq \Gamma'(m - t)(t - t') - [l(t) - l(t')] \\
&\geq G\frac{\Gamma(m - t)(t - t')}{m - t} - [l(t) - l(t')] \\
&\geq G\frac{\Gamma(m - t)(t - t')}{m - t} - G(t - t') > 0,
\end{aligned}$$

where the first inequality uses the convexity of Γ, the second inequality uses (6.1), and the very last strict inequality uses the fact that $\Psi_A(t,m) \leq 0$, so that $\Gamma(m-t) > m - l(t) > m - t$. The proof for Ψ_B uses an entirely analogous argument. $\square$

Lemma 6 and the fact that $\Psi_A(m,m) > 0$ imply that for each $m \in [0,1]$, either $\Psi_A(t,m) > 0$ for all $t \in [0,m]$, or there is a *unique* value of t — call it $t(m)$ — at which $\Psi_A(t(m),m) = 0$. In the former case, set $t(m) = 0$ to complete the definition of this function t, where $t(m) < m$ for all $m \in [0,1]$. In an analogous way, define a function μ, by $\mu(m) = 1$ if $\Psi_B(\mu,m) > 0$ for all $\mu \in [m,1]$, or otherwise as the unique solution to $\Psi_B(\mu,m) = 0$, so that $\mu(m) > m$, for all $m \in [0,1]$. Now define a mapping ζ as the conditional expectation $\zeta(m) = \mathbb{E}(x|x \in [t(m),\mu(m)])$. Clearly, ζ is continuous, and so has a fixed point. Pick any such fixed point; call it m^* and fix it for the rest of this proof. Define $(t^*,\mu^*) \equiv (t(m^*),\mu(m^*))$. This generates a random-order convention with the following properties:

(i) $0 \leq t^* < m^* < \mu^* \leq 1$; in particular, the randomization zone is always non-empty.

(ii) If $t^* > 0$, then $\Psi_A(t,m^*) < 0$ when $t < t^*$ and $\Psi_A(t,m^*) > 0$ when $t > t^*$, where $\Psi_A(t^*,m^*) = 0$, with analogous properties for μ^* and Ψ_B.

We claim that either at least one of the exception zones is nonempty. Suppose, on the contrary, that $t^* = 0$ and $\mu^* = 1$. Then $m^* = m_A$, and so $\Psi_A(0,m^*) = m_A + \frac{\delta}{2} - \Gamma(m_A) \geq 0$, because $t^* = 0$. But this contradicts (3.3).

In the particular case where f is symmetric, (6.1) need not hold. Set $m = 1/2$, and note that $\Psi_A(1/2,1/2) = 1/2 - l(1/2) + \delta/2 > 0$ but $\Psi_A(0,1/2) = 1/2 + \delta/2 - \Gamma(1/2) < 0$ by (3.4). Hence there exists $t^* > 0$ such that $\Psi_A(t^*,1/2) = 0$. Set $\mu^* = 1 - t^* < 1$. The symmetry here implies that $\Psi_B(\mu^*,1/2) = 0$ and that $m^* \equiv \mathbb{E}(x|x \in [t^*,\mu^*]) = \mathbb{E}(x|x \in [t^*,1-t^*) = 1/2$ so $(t^*,1-t^*)$ is our convention for the symmetric case, with all zones nontrivial.

We claim that any such solution described above — the fixed point for the non-symmetric case and the symmetric convention for the symmetric case — is an equilibrium random-order convention.

To prove the claim, consider first the range $0 \leq x < t^*$, presuming this range is non-empty. We show that in this range, Austen strictly prefers ρ to ⓡ, while Byron strictly prefers ρ to either ⓡ or α; the latter being only relevant when used by the convention; that is, when $\mu^* < 1$.

Begin with the claim for Austen. Observe that ρ yields Austen $l(t^*)-\Gamma(l(t^*)-x)$, while ⓡ yields Austen $m^*+\delta/2-\Gamma(m^*-x)$. The difference between the latter and the former is given by

$$\Lambda \equiv \left[m^* + \frac{\delta}{2} - \Gamma(m^* - x)\right] - \left[l(t^*) - \Gamma(l(t^*) - x)\right].$$

If $x \geq l(t^*)$, then $\Gamma(l(t^*) - x) = 0$ and so it is immediate that

$$\Lambda = [m^* - l(t^*)] + \frac{\delta}{2} - \Gamma(m^* - x) < [m^* - l(t^*)] + \frac{\delta}{2} - \Gamma(m^* - t^*) = \Psi_A(t^*, m^*) = 0,$$

where the last equality follows from the definition of t^* and the fact that $t^* > 0$ in this case.

If $x < l(t^*)$, then by the strict convexity of Γ (when positive), we see that $\Gamma(m^*-x)-\Gamma(l(t^*)-x) > \Gamma(m^* - t^*) - \Gamma(l(t^*) - t^*) = \Gamma(m^* - t^*)$, so that

$$\Lambda \equiv [m^* - l(t^*)] + \frac{\delta}{2} - [\Gamma(m^* - x) - \Gamma(l(t^*) - x)] < \Psi_A(t^*, m^*) = 0.$$

In short, $\Lambda < 0$ whenever $x < t^*$, establishing the claim for Austen.

Turning now to Byron's preferences in the range $x < t^*$, we first show that Byron strictly prefers ρ to ⓡ. Byron's lowest payoff under ρ occurs when $x = t^*$; it is $1 - l(t^*) + \delta - \Gamma(t^* - l(t^*))$. Under ⓡ, his payoff is $(1 - m^*) + (\delta/2) - \Gamma(x - m^*)$, which is bounded above by $(1 - m^*) + (\delta/2)$. Consequently, it suffices to show that

$$[m^* - l(t^*)] + \frac{\delta}{2} > \Gamma(t^* - l(t^*)). \tag{9.16}$$

Now, given that $t^* > 0$, we know that

$$\Psi_A(t^*, m^*) = [m^* - l(t^*)] + \frac{\delta}{2} - \Gamma(m^* - t^*) = 0, \tag{9.17}$$

Define $z_1 \equiv m^* - t^*$ and $z_2 \equiv m^* - l(t^*)$. Since $m^* \geq t^* \geq l(t)$ and (9.17) holds, the conditions of Lemma 3 are met. It follows that $m^* - t^* > t^* - l(t^*)$. Combining this inequality with (9.17) and recalling that Γ is increasing, we obtain (9.16).

To complete the proof of the claim when $x < t^*$, we show that Byron strictly prefers ρ to α. For ρ yields Byron a payoff of $1 - l(t^*) + \delta - \Gamma(x - l(t^*))$, whereas α yields at most $1 - h(\mu^*)$. The

difference between the two is then at least

$$[1 - l(t^*) + \delta - \Gamma(x - l(t^*))] - [1 - h(\mu^*)] = h(\mu^*) - l(t^*) + \delta - \Gamma(x - l(t^*))$$
$$> [m^* - l(t^*)] + \frac{\delta}{2} - \Gamma(m^* - l(t^*)) = 0,$$

establishing the claim.

We now apply this claim to verify that ρ is an equilibrium outcome in the range $x < t^*$. Note that ρ is the best choice for Byron, and ρ Pareto-dominates the default as both Austen and Byron strictly prefer it to ⓡ. Hence the claim follows from [I] and [II] of Section 3.6. Entirely analogous arguments apply to the range $x \in (\mu^*, 1]$ (when nonempty), where α is the equilibrium outcome.

Finally, consider the range $[t^*, \mu^*]$. Observe that if $x \in [t^*, m^*]$, Byron strictly prefers ⓡ to α. After all, ⓡ yields Byron a payoff of $(\delta/2)+(1-m^*)$, α yields $1-h(\mu^*)$, at most, and $h(\mu^*) > m^*$. By the construction of the threshold t^*, Austen strictly prefers ⓡ to ρ when $x \in (t^*, m^*]$. Hence ⓡ is the unique equilibrium outcome, when $x \in (t^*, m^*]$, by [I] of Section 3.6. (The zero-probability point $x = t^*$ can be attached either to outcome ⓡ or ρ.) By an analogous argument, when $x \in [m^*, \mu^*]$, Austen strictly prefers ⓡ to ρ. By the construction of the threshold μ^*, Byron strictly prefers ⓡ to α when $x \in [m^*, \mu^*)$. So the same argument holds for the subrange $[m^*, \mu^*]$, completing the proof of Theorem 3.

Proof of Theorem 4. As explained in the text, we only need to consider unplayed actions. These can only be α or ρ, as the central zone $[t^*, \mu^*]$ over which ⓡ is chosen is always nontrivial. Suppose, without loss of generality, that ρ is never played (the case in which α goes unplayed is completely parallel). Then $t^* = 0$, and so

$$\Psi_A(t, m^*) \equiv [m^* - l(t)] + \frac{\delta}{2} - \Gamma(m^* - t) > 0 \tag{9.18}$$

for all $t \in (0, m^*]$, with weak inequality at $t = 0$. Suppose, now, that a rational disruption ρ is observed off-path, and the public assigns to it the credit pair $(a, 1 - a)$ for some $a \in [0, 1]$. We need to show that no pair employing the disruption can have $x < m^*$.

In all cases, under ⓡ, the payoff to Austen is $m^* + (\delta/2) - \Gamma(m^* - x)$ and $(1 - m^*) + (\delta/2) - \Gamma(x - m^*)$ is the payoff to Byron. For ρ to invade it must be that both at least weakly prefer ρ to the default ⓡ, since, if one has the strict reverse preference, [s]he can veto ρ, by [I] of Section 3.6.

We must therefore have

$a-\Gamma(a-x) \geq m^*+(\delta/2)-\Gamma(m^*-x)$ and $(1-a)+\delta-\Gamma(x-a) \geq (1-m^*)+(\delta/2)-\Gamma(x-m^*)$.

Combining these two inequalities, we have

$$\Gamma(m^* - x) - \Gamma(a - x) \geq m^* + \frac{\delta}{2} - a \geq \Gamma(x - a) - \Gamma(x - m^*). \tag{9.19}$$

Suppose first that $a \leq m^*$. If $x \geq m^*$, the left-hand side of (9.19) must be zero, while the middle term is strictly positive, given that $m^* \geq a$, which is a contradiction. Therefore no pair with $x \geq m^*$ can deviate from ⓡ to ρ. No pair playing α would want to deviate to ρ either. For such a pair, $x \geq \mu^* > m^*$, and each co-author weakly prefers α to ⓡ, and even ⓡ cannot be weakly improved upon for both parties when $x \geq m^*$. Hence it must be that $x < m^*$, if $a \leq m^*$.

Suppose then that (9.19) holds for a nontrivial set of $x \geq 0$, and let $z \in (0, m^*]$ be the supremum of the values of x for which (9.19) holds. By the convexity of Γ, it is easy to see that the LHS of (9.19) is nonincreasing in x, while the RHS is nondecreasing. It follows that the set of deviants is given by the set $[0, z]$. By rationality of disruptions, it must be that

$$a = \mathbb{E}(x|x \leq z) = l(z). \tag{9.20}$$

In particular, $z > a$, so $\Gamma(a - z) = 0$. Therefore (9.19) implies

$$\Gamma(m^* - z) \geq m^* + \frac{\delta}{2} - a.$$

Using (9.20), we have

$$\Gamma(m^* - z) \geq m^* + (\delta/2) - l(z), \tag{9.21}$$

but (9.21) and $z > 0$ contradict (9.18). Hence there can be no nontrivial disruption when $a \leq m^*$.

This leaves the possibility that $a > m^*$. It cannot be that $a \in (m^*, m^* + \delta/2)$. For there has to be some pair with $x \geq a$ who would like to deviate. But for any $x \geq a$, the LHS of (9.19) is zero, while the middle term is positive, a contradiction. Therefore, $a \geq m^* + \delta/2$. Now the *lowest* x that might deviate is bounded below by m^*, for if $x < m^*$, then the LHS of (9.19) is negative, the middle term is non-positive, and the RHS is zero, a contradiction.

Thus we have established that for any pair $(x, 1-x)$ who might deviate to ρ, $x \geq m^*$. This completes the proof of the theorem.

Proof of Theorem 5. Consider first the total expected payoff under E_t, with threshold t as in Theorem 1. There are four relevant ranges for x:

If $x \in [0, t/2)$, then Austen's payoff is $t/2 - \Gamma(t/2 - x)$; whereas Byron's payoff is $1 - t/2 + \delta$.

If $x \in [t/2, t)$, Austen's payoff is $t/2$; whereas Byron's payoff is $1 - t/2 + \delta - \Gamma(x - t/2)$.

If $x \in [t, (1+t)/2)$, Austen's payoff is $(1+t)/2 + \delta - \Gamma((1+t)/2 - x)$; whereas Byron's payoff is $(1-t)/2$.

If $x \in [(1+t)/2, 1]$, Austen's payoff is $(1+t)/2 + \delta$; whereas Byron's payoff is $(1-t)/2 - \Gamma(x - (1+t)/2)$.

Hence the total expected payoff under E_t is given by

$$\begin{aligned} W(E) &\equiv 1 + \delta - \int_0^{t/2} \Gamma(t/2 - x)dx - \int_{t/2}^{t} \Gamma(x - t/2)dx \\ &\quad - \int_t^{(1+t)/2} \Gamma((1+t)/2 - x)dx - \int_{(1+t)/2}^{1} \Gamma(x - (1+t)/2)dx \\ &= 1 + \delta - 2\int_0^{t/2} \Gamma(x)dx - 2\int_0^{(1-t)/2} \Gamma(x)dx, \end{aligned} \tag{9.22}$$

where the second equality follows from a suitable change in variables.

For the ⓡ equilibrium, with thresholds at t^* and $1 - t^*$, there are three signals, with each signal range divided into two halves. Again, overall expected utility depends on the integral of the loss function over each of these ranges. An argument analogous to the one used to obtain (9.22) also applies to obtain the total expected payoff under ⓡ:

$$W(\text{ⓡ}) \equiv 1 + \delta - 4\int_0^{t^*/2} \Gamma(x)dx - 2\int_0^{1/2 - t^*} \Gamma(x)dx. \tag{9.23}$$

We must compare $W(\text{ⓡ})$ to $W(E)$. The following lemma will be useful:

Lemma 7. *If f is uniform,* (i) $t^* < 1/3$ *and* (ii) $t^* > t/2$.

Proof. (i) Define the function Ψ as $\Psi(t) = 1/2 + \delta/2 - t/2 - \Gamma(1/2 - t)$, which is the counterpart to (9.14) for this symmetric case. Clearly, $\Psi(1) > 0$ and by (3.3), $\Psi(0) < 0$. Since Γ is strictly

convex, it follows that t^* is the unique solution of $\Psi(t^*) = 0$. To show that $t^* < 1/3$ it then suffices to show that $\Psi(1/3) > 0$. However, $\Psi(1/3) = 1/3 + \delta/2 - \Gamma(1/6) > 0$, because $\Gamma(1/6) \leq 1/3$ given that $\Gamma(1/2) \leq 1$ from (3.4) and Γ is convex.

(ii) Suppose, on the contrary, that $t \geq 2t^*$. We have that $\Gamma(1/2 - t^*) + t^*/2 = 1/2 + \delta/2$. It follows that $\Gamma((1-t)/2) \leq \Gamma(1/2 - t^*) = 1/2 + \delta/2 - t^*/2 < 1/2 + \delta$, which contradicts (4.4). Therefore $t^* > t/2$, as claimed. □

With Lemma 7 in hand we complete the proof.

Define

$$D \equiv \left[\int_0^{t/2} \Gamma(x)dx + \int_0^{(1-t)/2} \Gamma(x)dx \right] - \left[2\int_0^{t^*/2} \Gamma(x)dx + \int_0^{1/2-t^*} \Gamma(x)dx \right].$$

Given (9.22) and (9.23), it suffices to show that

$$D > 0.$$

For clarity, we consider two cases. Suppose first that $t^* < t$. Then, using $t^* > t/2$ (Lemma 7), we see that

$$D = \int_{t^*/2}^{t/2} \Gamma(x)dx + \int_{1/2-t^*}^{1/2-t/2} \Gamma(x)dx - \int_0^{t^*/2} \Gamma(x)dx.$$

The total length of the intervals over which the positive integrals are taken is $t^*/2$, which is the length of the interval over which the negative integral is taken. In addition, $t^* < 1/3$ by Lemma 7, so the smallest value of $\Gamma(x)$ in the second integral — which is $\Gamma(1/2 - t^*)$ — is greater than the largest value of $\Gamma(x)$ from the negative integral, $\Gamma(t^*/2)$. It follows that $D > 0$ in this case.

Finally, suppose that $t^* \geq t$. In this case,

$$D = -\int_{t/2}^{t^*/2} \Gamma(x)dx + \int_{1/2-t^*}^{1/2-t/2} \Gamma(x)dx - \int_0^{t^*/2} \Gamma(x)dx,$$

again using $t^* > t/2$. The length of the interval over which the positive integral is taken is again equal to the combined length of the intervals over which the two negative integrals are taken. The smallest value of Γ in the positive integral, $\Gamma(1/2 - t^*)$, is greater than the largest value of Γ in either negative integral, which is $\Gamma(t^*/2)$, because $t^* < 1/3$. It follows that $D > 0$ yet again, completing the proof of the theorem. □