Epistemic Diversity and Editor Decisions: A Statistical Matthew Effect*

Remco Heesen† Jan-Willem Romeijn‡

October 9, 2017

Abstract

This paper offers a new angle on the common idea that the process of science does not support epistemic diversity. Under minimal assumptions on the nature of journal editing we prove that editorial procedures, despite being impartial in themselves, disadvantage less prominent research programs. In particular, we show that the quality of editorial decisions, as measured by false positives and negatives, is lower for programs that on the whole deliver fewer good papers. This purely statistical bias in article selection further skews the existing differences in the success rate and hence attractiveness of research programs, and exacerbates the reputation difference between the programs. Importantly, our results can be based on an assumption of real differences between the programs, but under certain circumstances our

*We thank audiences in Bristol, Hannover, and Bochum for valuable comments and discussion. RH was supported by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust.

†Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK. Email: rdh51@cam.ac.uk.

‡Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands. Email: j.w.romeijn@rug.nl.

results also hold when the programs are identical in terms of the quality of their output. The paper ends with a number of recommendations that may help promote scientific diversity through editorial decision making.

1 Introduction

The value of epistemic diversity in the sciences and the humanities has been argued extensively (e.g., Lakatos 1978, Longino 1990, Kitcher 1993). A scientific or scholarly discipline that harbors more research programs, i.e., a greater variety of methods and theories, will offer more balanced viewpoints and is better equipped to respond to challenges. In the words of Lakatos:

> The history of science has been and should be a history of competing research programmes (or, if you wish, 'paradigms'), but it has not been and must not become a succession of periods of normal science: the sooner competition starts, the better for progress. (Lakatos 1978, p. 69)

In the organization of science, we should therefore aim to facilitate diversity in research programs. This holds in particular for the peer review system with which all academic disciplines operate: a systematic bias towards a mono-culture is detrimental to scientific progress.

It is known that journal editors are prone to systematic (possibly unconscious) bias in favor of more prominent research programs. All the psychological and sociological factors that underlie a tendency of editorial decisions towards more prominent programs apply. Editors may suffer from a confirmation bias in assessing the quality of a research program, and they may choose conservatively among the available submissions with an eye on the reputation of the journal. Both with an eye on ethical considerations, and in view of the epistemic consequences of bias, journals are therefore well advised

to scrutinize their practices for biases of this kind. But unfortunately, these are not the only drivers of bias in editorial decisions.

This paper concerns biases that are rooted not in the prejudices of editors or reviewers, but rather in the statistical characteristics of editorial decision making. Our results confront two central notions in the review process: the probability that a paper gets accepted or rejected, and the average quality of accepted or rejected papers. Comparisons of different research programs with respect to these notions reveals that less well-established or otherwise vulnerable research programs are at a disproportional disadvantage. Hence, even if editors manage to purge their decision procedures of unconscious biases, they will be left with biases of a strictly statistical nature. These statistical biases contribute to the already existing tendency towards a monoculture in science: a purely statistical Matthew effect.

Our findings on editorial decisions rely on a number of modeling assumptions about the decision process: we presume that research papers have some latent inherent quality, that reviews offer a noisy measurement of this quality, and that editors base their decision to accept or reject a paper only on considerations of quality, informed by the reviews. The upshot of these assumptions is that the probability of acceptance is a monotonically increasing function of the latent quality. In what follows we take this notion of latent quality for granted but we will return to it in our discussion (section 4).

In the models we further assume that research programs have characteristics that can be spelled out in terms of probability distributions over the latent quality of papers. In particular, for our first result we assume that there is no quality difference between the programs. However, we imagine that the editor is more familiar with the researchers working within one research program than she is with another, and that as a consequence she has a more accurate estimation of the quality of their work. Under these assumptions, the first result, expounded in section 2, is that journal editors face a dilemma: either they accept more papers from the research program

with which they are more familiar, or the accepted papers from the more familiar program are on average of higher quality. If we add some additional assumptions editors fall prey to both.

So a reasonable editor, who does everything in her power to choose all and only high-quality papers for publication without regard for which research program produced it, will find herself accepting more papers from the research program that she is more familiar with. Moreover, looking back on her record she will typically feel justified in doing so, as the average quality of published papers that originated from her own research program has turned out to be higher. Assuming that editors are more likely to belong to established research programs, this makes it harder for new research programs to gain a foothold.

One possible response is that the editor abstains from using identifying author information, since the difference in familiarity with authors from the two programs is what drives the first result. But our second result, presented in section 3, shows that she will not rid herself of statistical Matthew effects. To arrive at this result, we assume that the programs differ in average latent quality. The result that we can then reach is similar to our first result. As may be expected, we can prove that more papers of the better program will get accepted. Less innocently, even though we acknowledge the difference between the programs by accepting more papers from the better program, on average those accepted papers will be of higher quality than the ones from the program with lower average quality. Moreover, among all the papers that meet a quality threshold, those from the better program are more likely to be accepted. In short, the reviewing process does a better job of picking out the high-quality papers from the better program.

Assuming that dominant or established research programs are likely to be better, at least initially, this makes it once again harder for less well-established programs to gain a foothold. Now perhaps this finding is a simple consequence of the fierce competition for space in respected scientific journals.

However, as we argue more extensively in the discussion, the mechanisms that benefit better, dominant, more established, or better-known programs merit careful study. As a preliminary motivation, consider the following:

> As long as a budding research programme can be rationally reconstructed as a progressive problemshift, it should be sheltered for a while from a powerful established rival. (Lakatos 1978, p. 71)

> [W]e sometimes want to maintain cognitive diversity even in instances where it would be reasonable for all to agree that one of two theories was inferior to its rival, and we may be grateful to the stubborn minority who continue to advocate problematic ideas. (Kitcher 1990, p. 7)

We can easily multiply other quotes that convey a similar preference for diversity in science, for example from Feyerabend (1975), Longino (1990), Hong and Page (2004), Zollman (2010) and Wylie (2014). Arguably, as pointed out by Philip Kitcher (personal communication), diversity in science may not be universally beneficial, partly because dissent may have adverse effects on the role of science in public discourse, and perhaps because some dissent moves beyond the confines of reasonable discussion. These caveats notwithstanding, we take the view that science benefits from the representation of a variety of perspectives to be fairly widely applicable, and we assume it throughout this paper. Given that we value and want to promote diversity of research programs, it is an open question whether and, if so, how the statistical biases of peer review should be counteracted.

Our current findings underline the challenges involved in safeguarding the diversity of research programs. We hope that they might also lead to reorienting our efforts in this respect. We do not suggest to cease critical assessment of our proneness to unconscious bias, but we warn that other sources of a tendency towards single-mindedness are at work. In the same vein, if a journal is seen to promote a dominant program to the detriment of

others, this cannot be ascribed simpliciter to biases at work in the editors. Instead, we should be aware that biases of a purely statistical nature may be at work in editorial decision making, and perhaps take steps to counteract these biases. In our discussion we will return to what these concrete steps might be.

2 Editor Decisions and the Value of Information

The results of this paper rely on a basic model of peer review. We imagine a scientific community with one journal, run by an editor who decides what gets published. The members of the community produce papers which they submit to the journal. Each paper has a quality q, measured by a real number (i.e., on a unidimensional scale). The editor aims to publish high-quality papers (in a way to be made precise momentarily). However, she faces uncertainty: the quality q is a latent variable, its value unknown to the editor. So when a paper arrives at the journal, all the editor has is a prior belief about its quality, in the form of a probability distribution over possible values of q.

The model thus adopts a common idea about peer review, namely that it is "the means by which one's equals assess the quality of one's scholarly work" (Eisenhart 2002, p. 241). Its aim is to guarantee "public confidence that high-quality academic work that makes a contribution to the accumulation of knowledge has been done" (Eisenhart 2002, p. 241). Conversely, bias in peer review may be defined as "any systematic effect on ratings unrelated to the true quality of the object being rated" (Blackburn and Hakel 2006, p. 378). These claims rely on a robust notion of quality, one on which it makes sense to speak of *the* quality of a journal submission.

When invoked so explicitly, the notion invites skepticism (cf. Lee et al. 2013). Rightfully so, we think, and we will return to this issue in section 4.

But normally these assumptions about quality are not made so explicit, and we take it that this picture of peer review is widely, if perhaps implicitly, shared among scientists. At any rate, whatever one thinks of the idea that journal submissions have a fixed, measurable quality, our models may be viewed as drawing out some of its consequences.

In her prior belief about a paper's quality, the editor takes into account the following factors. First, there are two competing research programs in this scientific community, the established research program H and the novel research program L, and each paper belongs to exactly one of these programs. Second, the editor is familiar with the work of some scientists in the community, but not others. The characteristics of particular scientists, insofar as the editor believes them to be relevant to the quality of their work, are represented in the model by a random variable K. If the editor knows a particular scientist k, she knows these characteristics ($K = k$) and takes them into account in her prior. If the author of a submission is unknown to the editor, she uses a generic prior that incorporates uncertainty about that scientist's characteristics.

	known author k	unknown author
research program H	$q \mid H, K = k$	$q \mid H$
research program L	$q \mid L, K = k$	$q \mid L$

Table 1: The prior distribution of q, given the research program the paper originates from and whether or not the author is known to the editor.

Based on these two factors (research program and author characteristics), submitted papers may be divided into four groups, with possibly different prior distributions (see table 1). But the editor actually believes both of these factors to be irrelevant, in the following sense. She believes that author characteristics follow the same distribution in the group of known scientists and in the group of unknown scientists. In other words, given only the fact that a scientist is known to the editor (but not the particular identity of the scientist), the editor expects to see the same distribution of quality as when

the scientist is unknown to the editor. In symbols (with $\sim$ denoting equality in distribution and $\mathbb{E}_K$ denoting expectation with respect to K):

$$\mathbb{E}_K[q \mid H, K] \sim q \mid H \quad \text{and} \quad \mathbb{E}_K[q \mid L, K] \sim q \mid L.$$

Moreover, the editor believes the distribution of quality to be the same for the two research programs (in symbols, $q \mid H \sim q \mid L$). In sum, the editor believes the papers from each of the four groups to be distributed over the quality values q in the same way.

Despite all this, knowing a particular scientist's characteristics may still be relevant. For example, suppose each of the four groups consisted of just two scientists, and in each group one of these scientists consistently produces high-quality work, the other low-quality. When the editor knows the individual scientists, she can take this into account. A reasonable decision procedure might be to accept all papers from the high-quality scientist and reject all papers from the low-quality scientist. But when she does not know the individual scientists, she cannot condition her decision procedure on author identity, and she might end up making worse decisions overall. This idea drives the main result of this section.

So far we have made no assumptions that distinguish the two research programs in any way. Here is the one difference we will assume: the editor knows the characteristics of a greater proportion of scientists in the established research program H than in the novel research program L. The idea behind this assumption is that the editor has had more time to familiarize herself with the key players in the more established program. Moreover, the editor herself is more likely to belong to this established program, making it even more probable that she knows a larger proportion of scientists working within it rather than within the novel program.

The editor solicits one or more reviews of the paper. The information gleaned from the reviewers' reports is summarized in a random variable R. We assume that R is independent of K given q, i.e., the quality of the pa-

per screens off any information about the author from the reviewers' report (perhaps because double-anonymous review is employed). Similarly, R is independent of the research program given q. Formally:

$$R \mid q \;\sim\; R \mid q, H \;\sim\; R \mid q, L \;\sim\; R \mid q, K = k.$$

The editor updates her belief about q based on the reviewers' report. Hence her posterior belief if she knows scientist k is either $q \mid R, H, K = k$ or $q \mid R, L, K = k$. If she does not know scientist k her posterior belief is $q \mid R, H$ or $q \mid R, L$. Now the editor has to make a decision D whether or not to accept the paper. We write $D \in \{A, \neg A\}$, where A denotes acceptance and $\neg A$ rejection. The editor aims to maximize the quality of accepted papers, i.e., her utility function is given by

$$u(D) = \begin{cases} q & \text{if } D = A, \\ q^* & \text{if } D = \neg A. \end{cases}$$

This says that if the editor accepts the paper her utility is equal to the real quality of the paper q, and if she rejects it her utility is some fixed constant value q^*.

As a result, the editor prefers to publish a paper if and only if $q > q^*$ (so the value of q^* indicates how selective the journal is). However, since the editor does not know the quality q she is facing a decision under uncertainty.

Being a rational editor, she maximizes her expected utility. (Note that if quality is measured in future citations, this is equivalent to maximizing the expected impact factor of the journal.) The expected utility of accepting the paper is the expected value of q, given her beliefs, i.e., it is equal to the mean of the editor's posterior distribution for the quality of the paper. The expected utility of rejecting the paper is simply q^*; no uncertainty there. So the editor accepts the paper if and only if the posterior mean quality exceeds q^*.

Given this model of editorial decisions and uncertainty, we are interested in two things. First, how likely is an arbitrary paper from an arbitrary scientist to be accepted, and how, if at all, does this depend on the research program the scientist belongs to? And second, what is the average quality of published papers originating from the two research programs?

Example 1. Suppose that quality follows a normal distribution with a mean that may be different for each author and a fixed known variance: $q \mid K = k \sim N(k, \sigma_q^2)$. Suppose further that author means are themselves normally distributed in the population: $K \sim N(\mu, \sigma_k^2)$. And suppose finally that the reviewer report provides a noisy but unbiased estimate of the quality of the paper, also with a normal distribution: $R \mid q \sim N(q, \sigma_r^2)$. If the overall acceptance rate of the journal is less than 50 % (or equivalently, if $q^* > \mu$) then the following inequalities hold (Heesen forthcoming):

$$\Pr(A \mid K) > \Pr(A) \quad \text{and} \quad \mathbb{E}[q \mid A, K] > \mathbb{E}[q \mid A].$$

If, as we have assumed, the editor knows a greater proportion of scientists from research program H than from research program L, it follows that

$$\Pr(A \mid H) > \Pr(A \mid L) \quad \text{and} \quad \mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L].$$

The results from this example are worrying. They show that in at least some circumstances, an editor who only aims to maximize the quality of accepted papers may show favoritism in the sense that she accepts papers from the established research program at a higher rate than those from the novel research program. Moreover, she will feel justified in doing so as the papers she accepts from program H are of higher quality (on average) than the papers she accepts from program L. But to the scientific community this is misleading because the distribution of quality is actually the same in the two research programs.

Are these results a peculiarity of the normal distributions assumed in

the example? The following theorem suggests an answer in the negative. It says that regardless of the distributions of q, K, and R, at least one of the inequalities from the example must hold.

Theorem 2. *If knowing the characteristics of a scientist sometimes makes a difference to the editor's decision (i.e., for some scientists there is a positive probability of getting a reviewer report such that the paper is accepted if the editor knows the scientist but rejected if the editor does not know the scientist, or vice versa), then*

$$\Pr(A \mid H) > \Pr(A \mid L) \quad \textit{or} \quad \mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L].$$

The proof is given in appendix A. It is based on the value of information theorem due to Good (1967). The idea is that the additional information the editor has available when she knows a scientist allows her to make better decisions.

The theorem shows that at least one of the following holds. Either the acceptance rate for papers from research program H is higher than the acceptance rate for papers from research program L, or accepted papers from program H are on average of higher quality than accepted papers from program L (even though the overall distribution of quality is the same in the two programs). We may formulate the result as a dilemma that the editor faces: either she will be seen to display a kind of favoritism by accepting papers from the established research program at a higher rate, or she will find that the papers she publishes from the established program turn out to be better papers (on average) than those she publishes from the novel program. In other words, the dilemma is between boosting research program H directly (by giving it more exposure) or indirectly (by creating the misleading impression that it produces higher quality work).

This is what we call a statistical Matthew effect: the established research program receives a boost despite its quality distribution being identical to

that of the novel research program, and despite the fact that neither the editor nor the reviewers are biased (all they do is measure quality). It is a Matthew effect (in the sense of Merton 1968) because the research program already enjoying a good reputation receives greater benefits when it delivers the same quality of work (as the novel research program). It is statistical because it arises from the underlying uncertainties in measuring quality as opposed to a specific preference from the editor or the reviewers.

3 Latent Differences

A salient feature of the model presented in the previous section is that the editor treats papers differently depending on their author: if the editor knows a particular scientist, she adjusts her prior based on what she knows about that scientist's characteristics, potentially lowering or raising the bar for publication. By and large, this means that scientists with a good track record will have their papers accepted even if the reviewers' report is merely lukewarm, whereas scientists with a poor track record need a glowing reviewers' report for acceptance.

In response to this we may prefer an editorial decision procedure that prevents this kind of differential treatment. We may want to rule out the use of prior information by the editor, effectively making editorial decisions depend only on the reviewers' report. This could be achieved by implementing triple-anonymous review.

Triple-anonymous review comes at a cost by giving up access to information that is potentially relevant for evaluating paper quality. This is suggested by our model—the editor does best in selecting for quality (at least by her own lights) if she follows the procedure of section 2, so deviating from that can only worsen her ability to select the best papers—and seems to be confirmed empirically by Laband and Piette (1994). Still, one may advocate triple-anonymous review for a number of reasons, either specifically to pre-

vent the statistical Matthew effect we described above, or to prevent various other types of biases (see Heesen forthcoming, Lee and Schunn 2010, p. 7).

Putting the other merits and demerits of triple-anonymous review to one side, here we suggest that it is not very successful as a response to the statistical Matthew effect. While triple-anonymous review (if implemented successfully) prevents the information asymmetry that drove our results above, in this section we argue similar phenomena can still occur if the quality distributions of the two research programs are different. So our claim will be that even a triple-anonymous reviewing procedure is vulnerable to a statistical Matthew effect under some circumstances.

We now present an adapted version of the model. As before, each paper has latent quality q. Papers belong to one of two research programs (H and L). The reviewers' report R provides information about the quality of the paper, and it does so in a way that is independent of the research program that the paper belongs to: $R \mid q, H \sim R \mid q, L$. We no longer distinguish between known and unknown authors as this plays no role here.

In the Bayesian model of the previous section, the relevance of the reviewers' report was in giving the editor relevant information to update from her prior beliefs to her posterior beliefs about the quality. In the model of this section, which is arguably more frequentist in character, the decision to accept a paper for publication is based exclusively on the reviewers' report. In particular, the paper is accepted if R exceeds a threshold value r^*, and any prior information about the quality of the paper is deemed irrelevant.

The reviewers' report R is a random variable which follows some probability distribution. We make no assumption on the shape of this distribution. We only assume that papers of higher quality have a greater chance of being accepted, in the following sense. Define the acceptance function a as the chance of acceptance given the latent quality q, i.e.,

$$a(q) := \Pr(A \mid q) = \Pr(R > r^* \mid q).$$

We assume that this function is (strictly) increasing, i.e., $q < q'$ implies $a(q) < a(q')$.

While we make essentially no assumptions on the distribution of R, in this section we do make some more substantial assumptions on the distribution of the latent. Let F_H be the distribution of quality for papers coming out of research program H, that is, F_H is the function such that $F_H(x) = \Pr(q \leq x \mid H)$. Similarly, let F_L be the quality distribution for research program L. If the quality distributions are continuous, as we will assume throughout, then the density functions f_H and f_L are defined as the derivatives of the distribution functions.

We make two key assumptions on the quality distributions, one on what they have in common and one on how they differ. First, we restrict our attention to continuous distributions whose density function is log-concave. A density function f is log-concave just in case

$$f(px + (1-p)y) \geq f(x)^p f(y)^{1-p}$$

for all $x, y \in \mathbb{R}$, and for all $p \in [0, 1]$. Log-concavity is a somewhat technical assumption restricting the shape of the distribution; among other things it entails that the distribution is unimodal. Log-concavity is satisfied by a range of well-known distributions, such as the normal, exponential, and uniform distributions. (The name comes from the fact that if f is log-concave, then the function $\log f$ is concave wherever it is well-defined.)

Second, we assume that the quality distributions for the two research programs have the same functional form, but that the average quality of papers produced by research program H is higher than the average of research program L. The idea is that the established program is able to reliably produce work of decent quality, whereas the novel program may suffer from startup problems. This assumption need of course not always be satisfied—in particular, the quality difference between research programs may point in the opposite direction—but here we explore cases where it holds. Hence,

in what follows we assume that work from the established program will be of higher quality, much like previously we assumed that scientists from the established program would be more familiar to the editor. We will discuss the reasonableness of this assumption in the next section.

More specifically and more formally, we assume the following. Let f be a log-concave density function supported on an interval $[b, c]$. (This means that $f(x) = 0$ if $x < b$ or $x > c$. We explicitly allow for the possibility that $b = -\infty$ and/or $c = \infty$.) Let F be the corresponding distribution function. Our assumption requires that there exist parameters μ_H and μ_L with $\mu_H > \mu_L$ such that

$$F_H(q) = F(q - \mu_H) \quad \text{and} \quad F_L(q) = F(q - \mu_L).$$

So we require that quality follows the same log-concave distribution in both research programs, differing only in that the distribution for research program H is shifted to the right compared to the distribution for research program L.

Our main result for this version of the model relates the probability that a paper is accepted for publication to the probability that its latent quality exceeds some threshold t. For interpreting the result, it is useful to think of the condition $q > t$ as asserting that the paper passes some threshold of suitability for publication. Given such a threshold, we may think of the goal of selecting for quality in terms of error rates: a false positive occurs when an unsuitable paper ($q \leq t$) is accepted for publication, and a false negative occurs when a suitable paper is rejected. The theorem says that regardless of the choice of threshold t, the peer review system works better for research program H in the sense that both the rates of false positives and false negatives are lower.

Theorem 3. *Let $t \in \mathbb{R}$ be any number in the support of f_H or f_L, i.e.,*

$b + \mu_L < t < c + \mu_H$. *Then*

$$\Pr(q > t \mid A, H) > \Pr(q > t \mid A, L) \quad \textit{and}$$
$$\Pr(A \mid q > t, H) \geq \Pr(A \mid q > t, L).$$

The latter inequality is strict unless the right tail of F is exponential, i.e., $f(q) \propto \exp\{-q\}$ for all $q > t + \mu_L$.

A proof of the theorem is given in appendix B. It generalizes results obtained in a different (psychometric) context by Borsboom et al. (2008). The intuition behind the proof is as follows. For any fixed quality q, the chance of acceptance does not depend on the research program, and the higher q, the higher the chance of acceptance. As a result, papers that are close to the suitability threshold t are at greatest risk of an error: those just above the threshold are less likely to be accepted than those far above it, and those just below the threshold are more likely to be accepted than those far below it. The distributional assumptions entail that among the suitable papers from research program L there are proportionally more papers close to the threshold than among suitable papers from research program H. This yields the result (which may be thought of as a continuous version of Simpson's paradox).

The peer review system we have modeled is "unbiased" in the following sense: papers of equal quality have the same chance of being accepted regardless of the research program they originated from. Perhaps surprisingly, such a peer review system may still be "biased" in the following sense: papers whose quality exceeds a threshold value may have different chances of acceptance depending on the research program they originated from. Under our assumptions, this phenomenon systematically favors the established research program.

Consider what this means for the novel research program. Of course, given its lower average quality, its overall acceptance rate is lower (see corollary 4

below). This is presumably as it should be. But the higher rate of false negatives means that on the rare occasion when the novel research program produces a good paper ($q > t$) it is relatively more likely to be rejected by the journal. And conversely, the higher rate of false positives means that when the novel research program manages to get a paper accepted for publication, it is relatively more likely to turn out to have been of low quality ($q \leq t$). Researchers forming an opinion of the novel program will quickly lose faith, pointing out that despite the editor's exclusive focus on latent quality, papers from the novel program are more often disappointing.

We have explained theorem 3 in terms of a notion of suitability for publication, which we have introduced somewhat artificially. But the theorem holds regardless of the choice of the threshold value t. So it follows from the theorem that the probability distribution of quality for those papers from research program H that get accepted for publication stochastically dominates the quality distribution for accepted papers from research program L: for any x, the probability that the quality of an accepted paper from program H is at least x is greater than the probability that the quality of an accepted paper from program L is at least x.

Again, researchers who see only what gets published will find that the novel research program consistently underperforms. Even among papers that appear in print, papers from the novel research program are consistently worse than papers from the established research program. As a result, researchers may even (falsely) suspect the editor of applying positive discrimination in favor of the novel research program: how else to explain the consistent difference in quality even among papers deemed publishable by the editor? Thus, we claim, there is a sense in which the peer review system seems biased against the novel research program even when we take into account the fact that its average quality is lower. Again this arises not from any individual bias at the level of the editor or the reviewers, but from the underlying probability distributions. This is the sense in which a statistical

Matthew effect operates in this second version of our model.

A few final remarks on this model. First, it may be helpful to phrase our result in a way that makes for a straightforward comparison with the results of the first model. Recall that, by theorem 2, either the acceptance rate or the average quality of published papers is higher for the established research program. Theorem 3 entails that both inequalities are satisfied in the present version of the model.

Corollary 4. *In the model of this section,*

$$\Pr(A \mid H) > \Pr(A \mid L) \quad \textit{and} \quad \mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L].$$

Moreover, the first inequality holds for any density function f, i.e., does not require the assumption of log-concavity.

Second, we may ask what happens when the distribution of quality differs between the two research programs in both mean and variance. We may again generalize (and slightly correct) results from Borsboom et al. (2008) to obtain a partial answer for the case where research program H has the greater variance. If we focus on predictive value rather than error, we get similar results for the case where research program L has the greater variance (cf. Borsboom et al. 2008, appendix C).

Theorem 5. *Define f and F as above. Let*

$$F_H(q) = F\left(\frac{q-\mu_H}{\sigma_H}\right) \quad \textit{and} \quad F_L(q) = F\left(\frac{q-\mu_L}{\sigma_L}\right).$$

Let $t \in \mathbb{R}$ be any number such that $\min\{\sigma_L b+\mu_L, \sigma_H b+\mu_H\} < t < \max\{\sigma_L c+\mu_L, \sigma_H c+\mu_H\}$. If

$$\sigma_H > \sigma_L \quad \textit{and} \quad \frac{\mu_H - t}{\sigma_H} \geq \frac{\mu_L - t}{\sigma_L}$$

then

$$\Pr(q > t \mid A, H) > \Pr(q > t \mid A, L) \quad \textit{and}$$
$$\Pr(A \mid q > t, H) > \Pr(A \mid q > t, L).$$

See appendix B for a proof.

Third, we note that our model in this section is less general than that of the previous section, requiring specific assumptions about the shape of the quality distribution (log-concavity) and the specific differences between the quality distributions of the two research programs (shifts in the mean and possibly variance). On the other hand it may also be considered more general, in that it allows for these differences between the programs, rather than assuming both programs to have identical distributions. We do not claim to have shown that a triple-anonymous reviewing procedure necessarily leads to a statistical Matthew effect; in particular, when the quality distributions are exactly equal it will not. But we have shown that in a nontrivial range of circumstances even triple-anonymous review is vulnerable to a statistical Matthew effect. This also has implications when our assumptions are not satisfied (e.g., when there is reason to believe quality distributions are not log-concave): in light of our work, it cannot simply be assumed that using triple-anonymous review suffices to prevent a statistical Matthew effect. One would rather have to show that the particular quality distribution at hand is not vulnerable to a statistical Matthew effect.

4 Discussion

The upshot of the foregoing is that editorial decision making is liable to biases of a purely statistical nature, and that these biases work against the diversity of research programs within a discipline. Firstly, since editors will be more familiar with authors from more established research programs, they

will more reliably choose good papers from those programs, thereby putting novel research programs at a disadvantage. Moreover, even if they choose to disregard this extra information and thus accept a decrease in the average quality of selected papers, this will not safeguard them against all statistical biases. A remaining problem is that the selection process has less favorable error rates for programs that have a lower average quality. Assuming that novel programs will have lower average quality, this bias again works against epistemic diversity.

Since we take such epistemic diversity to be beneficial, in the sense that it will enhance the versatility and thereby resilience of the discipline, these biases are detrimental. But before we turn to considering measures that will counteract or dissolve these biases, we critically assess the model in which we derived them. First we critically consider the assumption that papers have some latent quality. After that we discuss the relations among the quality of papers from a program, the familiarity of editors with the adherents of a program, and the program's novelty.

Can we claim that the model of editorial decision making is sufficiently similar to editorial practice, so that we are warranted to believe that the statistical biases indeed occur? The key feature of the model, we submit, is that it posits a latent quality of papers. But it is not at all clear that measuring paper quality is what editors and reviewers are actually doing, or what they should be doing.

Editorial practice consists in accepting and rejecting papers, and reviewer reports employ grades, allowing several degrees of acceptability, often accompanied by a narrative. The practice of any kind of peer review process—assessing papers for their "suitability" for publication—seems to implicitly commit editors and reviewers to some version of our story. A more extensive characterization of current editorial practices may well reveal that a latent, unidimensional, real-valued quality of papers is effectively induced by them, or at least an adequate representation of them. At the end of this discussion

we return to the possibility of doing away with the notion of quality. But we believe that the postulation of such a latent quality is an acceptable modeling assumption for current editorial practices.

Our reasons for thinking this are to do with a central phenomenon in psychometrics, namely the "positive manifold". Certainly there are many different aspects to the quality of a paper: the ingenuity of the experiment, the organization of the data, the clarity and statistical adequacy of the analysis, the coherence of the arguments, and so on. This is much the same for a psychological trait like cognitive ability, which involves processing speed, the size of working memory, pattern recognition, and the like. Spearman (1904) already noted that tests on all these aspects correlated positively, and that therefore we are warranted to postulate a single notion of general intelligence. We need not commit to its existence or causal efficacy (van der Maas et al. 2006), but we can include it as a modeling assumption. In much the same way, the various quality aspects of a paper will correlate positively, and it is therefore plausible that we can model paper quality as a single latent.

If we accept the notion in some form, it is still not clear what we should take to be expressed by the quality scale. Referees and editors rank papers according to several criteria, which are then compressed into a single grade or judgment. It is not clear what weighted combination of criteria is best taken as the quality (the familiar paradoxes of voting theory loom here). For our concerns, a particularly salient consideration is that one of these criteria might be the novelty or originality of the paper. That is, a paper may receive a high quality ranking because it brings a fresh perspective to a discipline, e.g., by working from within a new research program. If novelty by itself enhanced the quality of a paper, this would presumably undercut our main conclusions.

We do not deny that novelty of perspective is considered a virtue in a paper. At the level of a scientific discipline, epistemic diversity is a stand-alone virtue because it improves the versatility and hence resilience of the disci-

pline as a whole. However, when judging individual papers for inclusion in their own journal, editors will not normally factor such global considerations in. Their goal is to maintain their journal's status, and therefore to publish papers that offer good descriptions, reliable predictions, and convincing explanations. And these are in turn determined by the positive heuristics of the research program from which a paper originates, i.e., its core assumptions, and furthermore by the skill sets and the institutional and financial support of the researchers. At the level of the individual paper, novelty of perspective will not be a consideration in its own right, but at most one in a derivative sense, in that it may occasion benefits for the individual paper that matter to an editor (e.g., better descriptive, predictive or explanatory properties).

With the basic setup of the model in place, we can now turn to the more specific assumptions that drive the results of this paper. For the first result we assumed that the editor will be familiar with more researchers from an established program than with researchers from a new one, in the sense that she has a better view of the average quality of papers by researchers from this program. This seems to be fairly straightforward: a more established program will have had more exposure, and it is also more likely that the editor herself is associated with it. For the second result, however, the key assumption is that the average quality of papers from the more established program is higher. This assumption is far less straightforward, but we believe it can be motivated now that we have some more clarity on what latent quality is.

Recall that the descriptive, predictive and explanatory characteristics of a paper were supposed to be rooted in two sets of factors: the core assumptions of the adopted research program, and the skill and opportunity of the researchers. The latter set underpins the differences in the average quality of the programs: more established programs will have more social and monetary capital to make their research a success, they will have more developed methods and techniques, and also better training facilities. Additionally, the

program will be better equipped to recognize and signal quality and talent. If the core assumptions of the novel program are superior this may eventually come to light, but the idea is that the novel program starts at a disadvantage.

This concludes the critical assessment of the model. We have argued that existing editorial practices and the nature of research programs make the modeling assumptions plausible, at least in some contexts, so that the statistical biases that we have derived in the model indeed obtain. We devote the remainder of our paper to the question what we can do to counteract them.

One response to the first of the two results was already discussed at the start of section 3. This bias can be prevented by disallowing the information asymmetry required for the result. We could demand that no prior information about the author of a paper may be taken into account in evaluating it, analogously to the standing practice in criminal prosecution and psychological testing for the purpose of selection. One way to achieve this is by employing triple-anonymous review. However, this solution comes at a price. The editor knows that she foregoes information that would help her improve the average quality of papers in her journal.

Moreover, witness the second result, this approach fails to rule out all threats to epistemic diversity. We readily admit that the assumption of a lower average quality for the new program will not always hold but we believe we have motivated it sufficiently to say it holds in some contexts. One more-or-less direct repair of this can be constructed. As explained in the foregoing, the root cause of the differences in error rates is that the novel program has proportionally more papers that are near the quality cut-off point for inclusion in the journal. Accordingly, we can counteract the bias by directing more reviewer efforts towards papers that are borderline cases. To some extent this is already the standing practice, or so we think.

Nevertheless we would like to point to a problematic consequence of this repair. If more attention is given to papers near the quality cut-off point, then

this creates an asymmetry between the two programs of a different nature: the editorial office will effectively spend more of its reviewing resources on the novel program. In short, the repair again comes at a price, namely of creating a systematic advantage for the novel program (cf. the analogous discussion in Borsboom et al. 2008).

Some may find this an entirely acceptable price for safe-guarding epistemic diversity. Others may feel that the persistence of the problems with biases invites us to zoom out and search farther afield for a resolution of the problems. To this aim, we may return to the above critical assessment of our model. If the practice of editorial decision making is, at least in its basic setup, adequately represented by the model, then we can use this model to reconsider and subsequently change specific aspects of the practice. In particular, we might aim to eliminate the implicit adoption of a notion of quality, by rejecting the very idea that research papers need to go through a selection process.

What we are suggesting here is an altogether new way of looking at the system of science. Scientific publication is a process of collecting and curating papers, i.e., of regulated information sharing. Depending on what goals we take this information sharing to have, it may well turn out that it is better served by a system like ArXiv than by centralized collection and curation. The ultimate resolution of threats to epistemic diversity through biases in editorial decision making may be: to do away with editor decisions altogether. Depending on the details of such a system new problems will likely emerge, but there is reason to believe that statistical Matthew effects are not among them.

A Proof of the Value of Knowing the Author

Our proof relies on the following well-known result.

Theorem 6 (Good (1967))**.** *Given some choice problem, let D be a decision*

that maximizes expected utility relative to some prior beliefs and a utility function u. Let K be a random variable and let $D(K)$ be a decision that maximizes expected utility relative to the posterior beliefs (obtained from the prior beliefs by conditioning on the outcome of K) and utility u. Then

$$\mathbb{E}_K[\mathbb{E}[u(D(K))]] \geq \mathbb{E}[u(D)].$$

Moreover, the foregoing inequality is strict if there is a set of outcomes K_0 such that $\Pr(K \in K_0) > 0$ and $\mathbb{E}[u(D(k))] > \mathbb{E}[u(D)]$ for all $k \in K_0$ (i.e., decision D no longer maximizes expected utility if outcome k is observed).

In our model, we make a distinction between scientists known to the editor and scientists unknown to the editor, where knowing a scientist is represented as knowing the value of some random variable K that is potentially relevant to evaluating the quality of the scientist's paper. Let $D(K, R)$ be the decision taken by the editor if she knows the scientist's characteristics K and the reviewer report R and let $D(R)$ be the decision if the scientist is unknown, i.e., only the reviewer report R is known. Applying Good's theorem to our model yields the following.

Theorem 7. *Assume that there exists a set S of joint outcomes for K and R (i.e., members of S are pairs (k, r) where k is a possible outcome of K and r is a possible outcome of R) such that $D(k, r) \neq D(r)$ for all $(k, r) \in S$ and $\Pr((K, R) \in S) > 0$. Then*

$$\Pr(D(K, R) = A) > \Pr(D(R) = A) \quad \textit{or}$$
$$\mathbb{E}[q \mid D(K, R) = A] > \mathbb{E}[q \mid D(R) = A].$$

Proof. From theorem 6 we get that

$$\mathbb{E}_K[\mathbb{E}[u(D(K, R))]] \geq \mathbb{E}[u(D(R))],$$

with strict inequality if there is a set of outcomes for K with positive measure

such that $\mathbb{E}[u(D(k,R))] > \mathbb{E}[u(D(R))]$ for all k in that set. The theorem assumes that such sets of outcomes exist, so we have strict inequality in the above.

From the definition of u we know that $u(D(R)) = q$ if $D(R) = A$ and $u(D(R)) = q^*$ otherwise. Hence

$$\begin{aligned}\mathbb{E}[u(D(R))] &= \mathbb{E}[q \mid D(R) = A] \Pr(D(R) = A) + q^* \Pr(D(R) = \neg A) \\ &= q^* + \mathbb{E}[q - q^* \mid D(R) = A] \Pr(D(R) = A).\end{aligned}$$

Similarly

$$\begin{aligned}\mathbb{E}_K[\mathbb{E}[u(D(K,R))]] &= \mathbb{E}_K[\mathbb{E}[q \mid D(K,R) = A]] \Pr(D(K,R) = A) \\ &\quad + q^* \Pr(D(K,R) = \neg A) \\ &= q^* + \mathbb{E}_K[\mathbb{E}[q - q^* \mid D(K,R) = A]] \Pr(D(K,R) = A).\end{aligned}$$

The inequality obtained from theorem 6 entails

$$\begin{gathered}\mathbb{E}_K[\mathbb{E}[q - q^* \mid D(K,R) = A]] > \mathbb{E}[q - q^* \mid D(R) = A] \quad \text{or} \\ \Pr(D(K,R) = A) > \Pr(D(R) = A).\end{gathered}$$

Since q^* is a constant, the former inequality is equivalent to the one stated in the theorem. □

The above theorem assumes that there exists a set of outcomes S for K and R of positive probability such that $D(k,r) \neq D(r)$ for all $(k,r) \in S$. This is a more formally precise statement of the assumption made in theorem 2 that knowing the characteristics of a scientist sometimes makes a difference to the editor's decision. Theorem 2 follows as a corollary of theorem 7.

Proof of theorem 2. Conditional on whether or not the editor knows the characteristics of the scientist who wrote the paper, knowing which research program the paper belongs to is completely irrelevant: both the quality distri-

bution and the decision procedure used are identical for research programs H and L. It follows that both the probability of acceptance and the average quality of published papers are the same, i.e.,

$$\begin{aligned}
\Pr(D(K,R) = A \mid H) &= \Pr(D(K,R) = A \mid L),\\
\Pr(D(R) = A \mid H) &= \Pr(D(R) = A \mid L),\\
\mathbb{E}[q \mid D(K,R) = A, H] &= \mathbb{E}[q \mid D(K,R) = A, L],\\
\mathbb{E}[q \mid D(R) = A, H] &= \mathbb{E}[q \mid D(R) = A, L].
\end{aligned}$$

From theorem 7 we get that either the first of the above four lines is greater than the second, or the third line is greater than the fourth. Let p_{KH} denote the proportion of scientists in research program H known to the editor and let p_{KL} denote the proportion of scientists in research program L known to the editor. Then

$$\begin{aligned}
\Pr(A \mid H) &= p_{KH} \Pr(D(K,R) = A \mid H) + (1 - p_{KH}) \Pr(D(R) = A \mid H),\\
\mathbb{E}[q \mid A, H] &= p_{KH}\mathbb{E}[q \mid D(K,R) = A, H] + (1 - p_{KH})\mathbb{E}[q \mid D(R) = A, H],
\end{aligned}$$

and similarly for $\Pr(A \mid L)$ and $\mathbb{E}[q \mid A, L]$. The result follows from the assumption that $p_{KH} > p_{KL}$. □

B Proof of the Consequences of Latent Differences

For our purposes, the following characterization of log-concave densities is key. See Saumard and Wellner (2014, p. 97) for a proof.

Theorem 8. *Density function f is log-concave if and only if the family of densities f_G defined by $f_G(q) := f(q - \mu_G)$ has monotone likelihood ratios,*

i.e.,

$$\frac{f_H(q)}{f_L(q)} = \frac{f(q-\mu_H)}{f(q-\mu_L)} \geq \frac{f(q'-\mu_H)}{f(q'-\mu_L)} = \frac{f_H(q')}{f_L(q')}$$

whenever $q > q'$, $\mu_H > \mu_L$, $f_L(q) > 0$, *and* $f_L(q') > 0$.

The above theorem is used in the proof of our main result.

Proof of theorem 3. Let $f_H = F'_H$ and $f_L = F'_L$ be the density functions for the latent in the two groups. We first consider the distribution of quality conditional upon acceptance. Note that

$$\Pr(q > t \mid A, H) = \frac{\Pr(q > t, A \mid H)}{\Pr(A \mid H)} = \frac{\int_t^\infty a(q) f_H(q)\, dq}{\int_{-\infty}^\infty a(q) f_H(q)\, dq},$$
$$\Pr(q > t \mid A, L) = \frac{\int_t^\infty a(q) f_L(q)\, dq}{\int_{-\infty}^\infty a(q) f_L(q)\, dq}.$$

Consider the following special cases:

- If $c + \mu_L \leq t < c + \mu_H$ we are done immediately because $\Pr(q > t \mid A, H) > 0 = \Pr(q > t \mid A, L)$.
- If $b + \mu_L < t \leq b + \mu_H$ we are done immediately because $\Pr(q > t \mid A, H) = 1 > \Pr(q > t \mid A, L)$.
- If $c + \mu_L \leq b + \mu_H$ we are done immediately because for any value of t either $\Pr(q > t \mid A, H) = 1$ or $\Pr(q > t \mid A, L) = 0$.

So for the remainder of the proof we need only consider the case where $b + \mu_H < t < c + \mu_L$. By theorem 8 f_H/f_L is a non-decreasing function of q whenever it exists. This function exists for all q such that $f_L(q) > 0$, so in particular for $q \in (t, c + \mu_L)$. Thus

$$\begin{aligned}\Pr(q > t \mid A, H) &= \frac{\int_t^{c+\mu_L} a(q) f_H(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) f_H(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq} \\ &= \frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq + \int_{c+\mu_L}^{c+\mu_H} a(q) f_H(q)\, dq}.\end{aligned}$$

Since $b + \mu_H < t$, the numerator of this fraction is strictly smaller than the denominator, i.e., $\Pr(q > t \mid A, H) < 1$. It follows that

$$\Pr(q > t \mid A, H) \geq \frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq},$$

with strict inequality if $c < \infty$. Hence it suffices to show that

$$\frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_t^{c+\mu_L} a(q) f_L(q)\, dq} \geq \frac{\int_{b+\mu_L}^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_{b+\mu_L}^{c+\mu_L} a(q) f_L(q)\, dq},$$

with strict inequality if $c = \infty$. Let X be a random variable whose density function is given by

$$f_X(x) = \frac{a(x) f_L(x)}{\int_{b+\mu_L}^{c+\mu_L} a(u) f_L(u)\, du}$$

for all x. Then the above inequality is equivalent to

$$\mathbb{E}\left[\frac{f_H(X)}{f_L(X)} \mid X > t\right] \geq \mathbb{E}\left[\frac{f_H(X)}{f_L(X)}\right].$$

This inequality holds because f_H/f_L is non-decreasing by theorem 8. It remains to show that this inequality holds strictly if $c = \infty$. Equivalently, it remains to show that, if $c = \infty$,

$$\mathbb{E}\left[\frac{f_H(X)}{f_L(X)} \mid X > t\right] > \mathbb{E}\left[\frac{f_H(X)}{f_L(X)} \mid X \leq t\right].$$

Because f_H/f_L is non-decreasing, $f_H(t)/f_L(t)$ is a lower bound for the left-hand side of this inequality, and an upper bound for the right-hand side. Since t is assumed to be in the support of both f_H and f_L, $f_H(t)/f_L(t) > 0$.

If $b > -\infty$, then for $b + \mu_L < x < b + \mu_H$ we have $f_H(x) = 0$. Hence, conditional on $X < t$, $f_H(X)/f_L(X) = 0$ with positive probability, and

thus the expectation on the right-hand side must be strictly smaller than $f_H(t)/f_L(t)$.

If $b = -\infty$, then the inequality is strict unless $f_H(x)/f_L(x) = f_H(t)/f_L(t)$ for all $x \in \mathbb{R}$. But that happens only if $f_H = f_L$, i.e., if $F_H = F_L$. But we know that $F_H \neq F_L$ because these distributions are obtained from F by adding different constants $\mu_H \neq \mu_L$.

This concludes the proof for the distribution of quality given acceptance. Now consider the probability of acceptance given $q > t$.

$$\begin{aligned}\Pr(A \mid q > t, H) &= \frac{\Pr(q > t, A \mid H)}{\Pr(q > t \mid H)} = \frac{\int_t^\infty a(q) f_H(q)\, dq}{\int_t^\infty f_H(q)\, dq} \\ &= \mathbb{E}[a(q) \mid q > t, H], \\ \Pr(A \mid q > t, L) &= \frac{\int_t^\infty a(q) f_L(q)\, dq}{\int_t^\infty f_L(q)\, dq}.\end{aligned}$$

Note that if $c + \mu_L \leq t < c + \mu_H$ then $\Pr(q > t \mid L) = 0$. This would mean that $\Pr(A \mid q > t, L)$ is not defined, so we set this case aside and suppose that $t < c + \mu_L$.

We may write $\mathbb{E}[a(q) \mid q > t, H]$ as a weighted average of $\mathbb{E}[a(q) \mid q > c + \mu_L, H]$ and $\mathbb{E}[a(q) \mid t < q \leq c + \mu_L, H]$. Since a is an increasing function,

$$\mathbb{E}[a(q) \mid q > c + \mu_L, H] > a(c + \mu_L) > \mathbb{E}[a(q) \mid t < q \leq c + \mu_L, H].$$

Hence

$$\begin{aligned}\Pr(A \mid q > t, H) &\geq \mathbb{E}[a(q) \mid t < q \leq c + \mu_L, H] \\ &= \frac{\int_t^{c+\mu_L} a(q) f_H(q)\, dq}{\int_t^{c+\mu_L} f_H(q)\, dq} = \frac{\int_t^{c+\mu_L} a(q) \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_t^{c+\mu_L} \frac{f_H(q)}{f_L(q)} f_L(q)\, dq},\end{aligned}$$

with strict inequality if $c < \infty$. Then it suffices to show that

$$\mathbb{E}\left[\frac{f_H(Y)}{f_L(Y)}\right] = \frac{\int_t^{c+\mu_L} \frac{f_H(q)}{f_L(q)} a(q) f_L(q)\, dq}{\int_t^{c+\mu_L} a(q) f_L(q)\, dq} \geq \frac{\int_t^{c+\mu_L} \frac{f_H(q)}{f_L(q)} f_L(q)\, dq}{\int_t^{c+\mu_L} f_L(q)\, dq} = \mathbb{E}\left[\frac{f_H(Z)}{f_L(Z)}\right],$$

where Y and Z's density functions are given respectively by

$$f_Y(x) = \frac{a(x) f_L(x)}{\int_t^{c+\mu_L} a(u) f_L(u)\, du} \quad \text{and} \quad f_Z(x) = \frac{f_L(x)}{\int_t^{c+\mu_L} f_L(u)\, du}$$

if $x > t$ ($f_Y(x) = f_Z(x) = 0$ otherwise). Note that whenever $x > t$,

$$\frac{f_Y(x)}{f_Z(x)} \propto a(x),$$

which is increasing in x by assumption. So Y has relatively higher density for high values. Since, moreover, f_H/f_L is non-decreasing, it follows that $\mathbb{E}[f_H(Y)/f_L(Y)] \geq \mathbb{E}[f_H(Z)/f_L(Z)]$.

The inequality is an equality only if $f_H(q)/f_L(q) = f_H(t)/f_L(t)$ for all $q > t$. This happens if and only if $f(q) \propto \exp\{-q\}$. □

Proof of corollary 4. By theorem 3, we have

$$\Pr(q > t \mid A, H) > \Pr(q > t \mid A, L)$$

for all $t \in (b+\mu_L, c+\mu_H)$. For any t outside this interval, the above inequality is an equality (both probabilities are one if $t \leq b + \mu_L$ and zero otherwise). Thus the distribution $q \mid A, H$ first-order stochastically dominates the distribution of $q \mid A, L$. It follows that $\mathbb{E}[q \mid A, H] > \mathbb{E}[q \mid A, L]$.

We could use the other inequality from theorem 3 to establish the inequality in acceptance rates, but then we would need to worry about the special case where the right tail of f is exponential. Instead we provide a simple direct proof of the inequality in acceptance rates, which also shows

that the assumption that f is log-concave is superfluous.

$$\begin{aligned}\Pr(A \mid H) &= \int_{b+\mu_H}^{c+\mu_H} a(q) f(q-\mu_H)\, dq \\ &= \int_{b+\mu_L}^{c+\mu_L} a(u+\mu_H-\mu_L) f(u-\mu_L)\, du \\ &> \int_{b+\mu_L}^{c+\mu_L} a(u) f(u-\mu_L)\, du = \Pr(A \mid L). \end{aligned}$$ □

Proof of theorem 5. By the chain rule, F_H and F_L are differentiable and their densities are given by

$$f_H(q) = \frac{1}{\sigma_H} f\left(\frac{q-\mu_H}{\sigma_H}\right) \quad \text{and} \quad f_L(q) = \frac{1}{\sigma_L} f\left(\frac{q-\mu_L}{\sigma_L}\right)$$

for all q.

Consider the probability that a paper from research program H is accepted and its quality q exceeds t. Using the substitution $q = t + \frac{\sigma_H}{\sigma_L}(u-t)$ we find:

$$\begin{aligned}&\Pr(q > t, A \mid H) = \int_t^{\sigma_H c+\mu_H} a(q) \frac{1}{\sigma_H} f\left(\frac{q-\mu_H}{\sigma_H}\right) dq \\ &= \int_t^{\sigma_L c + t + \frac{\sigma_L}{\sigma_H}(\mu_H - t)} a\left(t + \frac{\sigma_H}{\sigma_L}(u-t)\right) \frac{1}{\sigma_L} f\left(\frac{1}{\sigma_L}\left(u - t - \frac{\sigma_L}{\sigma_H}(\mu_H - t)\right)\right) du \\ &= \int_t^{\sigma_L c+\mu'} a\left(t + \frac{\sigma_H}{\sigma_L}(u-t)\right) g(u-\mu')\, du,\end{aligned}$$

where g is the function given by $g(x) = f(x/\sigma_L)/\sigma_L$ and $\mu' = t + \frac{\sigma_L}{\sigma_H}(\mu_H - t)$. Since $u > t$ and $\sigma_H > \sigma_L$ we have $t + \frac{\sigma_H}{\sigma_L}(u-t) > u$. Using the fact that a is an increasing function:

$$\Pr(q > t, A \mid H) > \int_t^{\sigma_L c+\mu'} a(u) g(u-\mu')\, du.$$

Analogously, we find that

$$\Pr(q \leq t, A \mid H) < \int_{\sigma_L b+\mu'}^{t} a(u)g(u-\mu')\,du.$$

Applying these two inequalities yields

$$\begin{aligned}\Pr(q > t \mid A, H) &= \frac{\Pr(q > t, A \mid H)}{\Pr(q > t, A \mid H) + \Pr(q \leq t, A \mid H)} \\ &> \frac{\int_t^{\sigma_L c+\mu'} a(u)g(u-\mu')\,du}{\int_{\sigma_L b+\mu'}^{\sigma_L c+\mu'} a(u)g(u-\mu')\,du}.\end{aligned}$$

Note that the function g is itself a density function: in particular, if f is the density function of some random variable X, then g is the density function of the random variable $\sigma_L X$. Since f is log-concave, and log-concavity is preserved by affine transformations (Saumard and Wellner 2014, p. 57), g is also log-concave.

But then we can apply theorem 3! In particular, the condition $(\mu_H - t)/\sigma_H \geq (\mu_L - t)/\sigma_L$ is equivalent to $\mu' \geq \mu_L$. Hence by theorem 3:

$$\begin{aligned}\Pr(q > t \mid A, H) &> \frac{\int_t^{\sigma_L c+\mu'} a(u)g(u-\mu')\,du}{\int_{\sigma_L b+\mu'}^{\sigma_L c+\mu'} a(u)g(u-\mu')\,du} \\ &\geq \frac{\int_t^{\sigma_L c+\mu_L} a(u)g(u-\mu_L)\,du}{\int_{\sigma_L b+\mu_L}^{\sigma_L c+\mu_L} a(u)g(u-\mu_L)\,du} \\ &= \frac{\int_t^{\sigma_L c+\mu_L} a(u)\frac{1}{\sigma_L}f\left(\frac{u-\mu_L}{\sigma_L}\right)\,du}{\int_{\sigma_L b+\mu_L}^{\sigma_L c+\mu_L} a(u)\frac{1}{\sigma_L}f\left(\frac{u-\mu_L}{\sigma_L}\right)\,du} = \Pr(q > t \mid A, L).\end{aligned}$$

This proves the first inequality. The second inequality is quite similar. Consider the probability that the quality of a paper from research program H exceeds t. Using again the substitution $q = t + \frac{\sigma_H}{\sigma_L}(u-t)$ we find:

$$\Pr(q > t \mid H) = \int_t^{\sigma_H c+\mu_H} \frac{1}{\sigma_H} f\left(\frac{q-\mu_H}{\sigma_H}\right)\,dq = \int_t^{\sigma_L c+\mu'} g(u-\mu')\,du.$$

Combining this with the result for $\Pr(q > t, A \mid H)$ from the first half of the proof, we see that

$$\Pr(A \mid q > t, H) = \frac{\Pr(q > t, A \mid H)}{\Pr(q > t \mid H)} > \frac{\int_t^{\sigma_L c + \mu'} a(u) g(u - \mu')\, du}{\int_t^{\sigma_L c + \mu'} g(u - \mu')\, du}.$$

Since g is log-concave and $\mu' \geq \mu_L$, we can apply theorem 3 to get

$$\begin{aligned}
\Pr(A \mid q > t, H) &> \frac{\int_t^{\sigma_L c + \mu'} a(u) g(u - \mu')\, du}{\int_t^{\sigma_L c + \mu'} g(u - \mu')\, du} \\
&\geq \frac{\int_t^{\sigma_L c + \mu_L} a(u) g(u - \mu_L)\, du}{\int_t^{\sigma_L c + \mu_L} g(u - \mu_L)\, du} \\
&= \frac{\int_t^{\sigma_L c + \mu_L} a(u) \frac{1}{\sigma_L} f\left(\frac{u - \mu_L}{\sigma_L}\right) du}{\int_t^{\sigma_L c + \mu_L} \frac{1}{\sigma_L} f\left(\frac{u - \mu_L}{\sigma_L}\right) du} = \Pr(A \mid q > t, L). \qquad \square
\end{aligned}$$

References

Jessica L. Blackburn and Milton D. Hakel. An examination of sources of peer-review bias. *Psychological Science*, 17(5):378–382, 2006. doi: 10.1111/j.1467-9280.2006.01715.x. URL `http://dx.doi.org/10.1111/j.1467-9280.2006.01715.x`.

Denny Borsboom, Jan-Willem Romeijn, and Jelte M. Wicherts. Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, 13(2):75–98, Jun 2008. doi: 10.1037/1082-989X.13.2.75. URL `http://dx.doi.org/10.1037/1082-989X.13.2.75`.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. ISSN 1573-1898. doi: 10.1023/A:1016082229411. URL `http://dx.doi.org/10.1023/A:1016082229411`.

Paul Feyerabend. *Against Method*. New Left Books, London, 1975.

I. J. Good. On the principle of total evidence. *The British Journal for the Philosophy of Science*, 17(4):319–321, 1967. ISSN 00070882. URL http://www.jstor.org/stable/686773.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, forthcoming. doi: 10.1007/s11098-017-0895-4. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46): 16385–16389, 2004. doi: 10.1073/pnas.0403723101. URL http://www.pnas.org/content/101/46/16385.abstract.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. ISSN 0022362X. URL http://www.jstor.org/stable/2026796.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993. ISBN 0195046285.

David N. Laband and Michael J. Piette. Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1):194–203, 1994. ISSN 00223808. URL http://www.jstor.org/stable/2138799.

Imre Lakatos. *The Methodology of Scientific Research Programmes*. Cambridge University Press, Cambridge, 1978.

Carole J. Lee and Christian D. Schunn. Philosophy journal practices and opportunities for bias. *American Philosophical Association Newsletter on Feminism and Philosophy*, 10(1):5–10, 2010. ISSN 2155-9708. URL http://www.apaonline.org/?feminism_newsletter.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Helen E. Longino. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, 1990. ISBN 9780691020518.

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968. ISSN 00368075. URL http://www.jstor.org/stable/1723414.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114, 2014. doi: 10.1214/14-SS107. URL http://dx.doi.org/10.1214/14-SS107.

C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. ISSN 00029556. doi: 10.2307/1412107. URL http://www.jstor.org/stable/1412107.

Han L. J. van der Maas, Conor V. Dolan, Raoul P. P. P. Grasman, Jelte M. Wicherts, Hilde M. Huizenga, and Maartje E. J. Raijmakers. A dynamic model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4):842–861, 2006. doi: 10.1037/0033-295X.113.4.842. URL http://dx.doi.org/10.1037/0033-295X.113.4.842.

Alison Wylie. Community-based collaborative archaeology. In Nancy Cartwright and Eleonora Montuschi, editors, *Philosophy of Social Science: A New Introduction*, chapter 4, pages 68–82. Oxford University Press, Oxford, 2014.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*,

72(1):17–35, 2010. ISSN 01650106. URL http://www.jstor.org/stable/20642278.