INTERDISCIPLINARITY CAN AID THE SPREAD OF BETTER METHODS BETWEEN SCIENTIFIC COMMUNITIES

PAUL E. SMALDINO[1,*] AND CAILIN O'CONNOR[2]

November 10, 2020

Abstract: Why do bad methods persist in some academic disciplines, even when they have been clearly rejected in others? What factors allow good methodological advances to spread across disciplines? In this paper, we investigate some key features determining the success and failure of methodological spread between the sciences. We introduce a model that considers factors like methodological competence and reviewer bias towards one's own methods. We show how self-preferential biases can protect poor methodology within scientific communities, and lack of reviewer competence can contribute to failures to adopt better methods. We further argue, however, that input from outside disciplines, especially in the form of peer review and other credit assignment mechanisms, can help break down barriers to methodological improvement.

1. Introduction

Dr. Pants is an established scientist and a well-known expert in her field. One day, she is asked to review a paper for a highly ranked journal. The paper asks questions similar to the ones she investigates, but uses an unfamiliar method borrowed from another field. The authors claim that the method used by Dr. Pants and her colleagues is flawed and that their own method is more accurate and less error-prone. Approval from Dr. Pants would help increase the visibility of the new method and the prominence of the paper's authors. While the authors are correct, Dr. Pants is biased in favor of the methods used in her own previous work. Furthermore, having never employed this new method herself, she is not fully competent to evaluate its relative quality. She decides not to risk it, and recommends rejection.

There are many real cases where select scientific communities have continued to use poor methods, even after these methods were critiqued or abandoned in other areas of science and better methods have become available. The replication crisis of the last decade has shed light, for instance, on the long term use of problematic practices in areas like social psychology and biomedical research. And even as some scientific disciplines adopt reforms like open data, preregistration, and commitment to more rigorous

[1]University of California, Merced
[2]University of California, Irvine
E-mail address: psmaldino@ucmerced.edu.

statistical training, others remain resistant.[1] On the face of it, this pattern seems surprising. We expect scientific communities to share the normative goal of seeking truth, and thus expect that methods promoting this goal should be widely adopted. What is going wrong? Why do problematic methods persist in some communities, even after they have been clearly rejected in others? And what solutions can we identify?

We develop a model intended to investigate some key features of scientific communities that may contribute to the failure of disciplines to adopt better methods. Our model considers groups of scientists who choose between methods that are more or less likely to yield epistemic successes—that is, to contribute new knowledge or understanding of some system or phenomenon. The research produced by these scientists is judged by peers, and methods that lead to successful publications tend to be adopted by others in the field. When review tracks only epistemic success, there is no problem. Communities adopt superior methods. We are interested in cases where reviewers are either incompetent to judge between methods, or else show biases for the (possibly poor) methods already used by themselves and their peers. We find that these factors can lead to the stable persistence of poor methods. This result echoes and expands upon previous work by Akerlof and Michaillat (2018), who focus on the role of this sort of self-preferential bias in preventing the spread of superior paradigms.

If bias for existing methods is strong, and competence to assess the quality of new methods is weak, are some scientific disciplines doomed to remain plagued with low quality methods? We argue that interdisciplinarity can help. When reviewers from disciplines with different competences and biases are able to judge work within a problematic discipline, and assign credit based on their own standards, better methods can spread within the original discipline. This finding bears upon claims from scholars such as Longino (1990) and Oreskes (2019), who have argued that diverse communities improve scientific reliability, partly because outsiders are better able to criticize existing paradigms when they do not share their underlying assumptions. But the long-term maintenance of diversity in a cooperative community is difficult due to the forces of conformity and other normative pressures (Axelrod, 1997; Henrich and Boyd, 1998; Smaldino and Epstein, 2015; Weatherall and O'Connor, 2020). Disciplinary structure can preserve a diversity of assumptions and competencies, while interdisciplinarity facilitates the flow of good research practices across the sciences.

The remainder of the paper will proceed as follows. In section 2 we discuss a case study which illustrates the importance of competence, bias, and interdisciplinarity in determining methodological practice. In section 3 we give some relevant background for our model, and discuss previous research looking at the failure of good methods to spread in science. Section 4 outlines our formal model. We present the results of our analysis in the subsequent two sections, first focusing on how methods spread in an isolated community (section 5), and then on interacting communities (section 6). We

[1] There is occasionally disagreement regarding the benefit of such practices, such as with preregistration (Szollosi et al., 2019). Such disagreement supports our perspective on the challenges in evaluating the relative benefits of adopting novel methods.

conclude in section 7 by discussing implications for the optimal structure of scientific communities.

2. MBI and the persistence of poor methodology in sports science

In this section we discuss a case study from the field of sports science.[2] Two factors we explore in our models—competence and self-preferential bias—seem to have played an important role in this case. And furthermore interdisciplinary feedback now seems to be playing a key role in disciplinary reform. To be clear, this is far from the only case where problematic methods have persisted in scientific disciplines, but it is a particularly illustrative one. In the conclusion we will briefly discuss a few further examples, as well as some cases that are less well addressed by the models presented here.

Foam rolling involves lying down, face up, with a soft polymer cylinder under one's back and rolling over it with the aim of relaxing muscles. How effective is foam rolling as a treatment for muscle ailments? MacDonald et al. (2014) examined the usefulness of foam rolling and purported to show that it reduces muscle soreness, improves range of motion, and even improves performance in activities like vertical jump height when used after exercise. Since publication in 2014, their paper has been cited nearly 300 times according to Google Scholar. But the paper depends on a statistical method called measurement based inference (MBI), which statisticians have widely criticized as misleading and incoherent, producing high rates of false positive findings (Sainani, 2018). A re-examination of the data from the foam rolling study using more rigorous statistical testing casts doubts on many of their claims (Lohse et al., 2020). Over the years, hundreds of published papers in the discipline of sports science have used MBI, and the unsoundness of the method throws many of these results into doubt (Lohse et al., 2020).

The method was introduced in a 2006 paper by the sports scientists Alan Batterham and Will Hopkins (Batterham and Hopkins, 2006) with the intent of supplanting standard frequentist statistical practices. Briefly, MBI involves comparing the risk that an intervention causes harm with the chance that it benefits an athlete. Under sufficiently low risk and sufficiently high chance of benefit, the method will proclaim a finding "substantial". Rather than providing equations or describing algorithms that detail the workings of MBI, the authors developed and distributed Microsoft Excel spreadsheets that allowed readers to implement their method without understanding it. Users could easily input data from an experiment, and the spreadsheets would output judgments related to harm and benefit.

Several factors seem to have contributed to the uptake of MBI in sports science. First, the authors, both prominent community members, actively promoted the method through a website run by Hopkins, and at sports science conferences.[3] Second, sports scientists, like scientists in many disciplines, tend to be relatively unfamiliar with the

[2]This case study draws heavily on the work of the science journalist Christie Aschwanden (Aschwanden and Nguyen, 2018; Aschwanden, 2018, 2019).

[3]As of the time of writing, Hopkins continues to actively promote the use of MBI, though Batterham seems to have distanced himself from it.

mathematical details of their statistical methods, and thus are unable to adjudicate the quality of such methods themselves (Vigotsky et al., 2020). Third, the ease of use and ready availability of the spreadsheets mentioned above seems to have been attractive to users. And last, the method has a high false positive rate. This means that researchers in sports science, who are often looking for small effects in small samples, were able to publish results using MBI that would otherwise have been statistically insignificant.[4]

Shortly after its introduction, statisticians Barker and Schofield (2008) pointed out flaws in the MBI method, and suggested a move to standard Bayesian statistics. But these suggestions were not taken up. And for most of a decade the use of MBI continued in sports science with little resistance. In Welsh and Knight (2015), another pair of statisticians raised further concerns about MBI, but these were again dismissed by many in the field. More recently another statistician, Kristin Sainani, pointed out that several claims by the original authors about MBI were incoherent. She also showed how the false positive rate of the method is unacceptably high, especially for studies with small samples sizes (exactly the sort of studies in which it is typically used) (Sainani, 2018; Lohse et al., 2020). Other statisticians have expressed surprise and skepticism upon learning about MBI. Biostatistician Andrew Vickers has described it as "a math trick that bears no relation to the real world" (Aschwanden and Nguyen, 2018).

In light of these critiques, there has recently been some movement to suppress the method in sports science, including a ban by the well-respected journal *Medicine and Science in Sports and Exercise* (MSSE). Responding in part to the use of MBI, a research team including both statisticians and sports scientists have called for better statistical training in sports science and for more collaboration with trained statisticians (Sainani et al., 2020). But despite this turn, at the time of this writing numerous articles continue to be published using variations of MBI.

There are a few things to highlight about this case. First, for the duration of its use in sport science, there has been no controversy or confusion among statisticians about the status of MBI—they have universally labeled it as a misleading methodology. Nevertheless, reviewers *within* sports science, in judging the research of their peers, have not rejected that research as faulty because of its reliance on MBI. This seems to be in part due of a lack of competence by many reviewers to assess statistical methods. It also seems to be due to the strength of disciplinary norms. Once the use of MBI became widespread, researchers who employed it themselves were perfectly willing to accept other papers using it for publication. Biostatistician Doug Everett, who wrote a commissioned editorial on MBI, says about it, "I almost get the sense that this is a cult. The method has a loyal following in the sports and exercise science community, but that's the only place that's adopted it" (Aschwanden and Nguyen, 2018).

A second thing to highlight, which will become relevant to the modeling results we discuss later on, is the role of statisticians in the slow move towards rejecting MBI. This includes their role as reviewers in the field. For example, the originators of MBI, Batterham and Hopkins, tried to publish a defense of the method in the journal *MSSE*.

[4]For work arguing that pressure to publish can lead to the widespread adoption of poor statistical methods, see Smaldino and McElreath (2016).

Reviewers well-versed in statistics rejected the paper. Hopkins and Batterham then submitted to the journal *Sports Medicine*, where reviewers from sports science, who Hopkins describes as having been "groomed" by him, accepted the paper (Aschwanden and Nguyen, 2018; Hopkins and Batterham, 2016). In other words, there was resistance from without, including during the reviewing process, even while many insiders happily accepted MBI. This resistance seems to have been crucial in the current turn against the method.

3. Bias, Competence, and Interdisciplinary Contact

In the next sections we will try to systematically investigate some of the key features at work in the MBI cases, and to test their relevance to the persistence of poor methodology. Before doing so, let us first address these key features—bias, competence, and interdisciplinary contact—at more length. We will follow this with a discussion of previous modeling literature.

3.1. **Bias and Competence.**

One focus of our investigation here is researcher *bias* towards one's own methods. Many communities have norms both for researchers to use the standard methods of the field and to publish in the discipline's top journals, the gatekeepers of which help to standardize the field's practices. Novel methods must often overcome conservative biases for these existing methods, a point noted by Kuhn (1962) and others. Recently, concerns have been raised about the adverse effects of conservative norms in rejecting both novel methodologies and novel questions in the social sciences (Akerlof, 2020; Barrett, 2020a; Stanford, 2019).

In addition, previous empirical work has found that scientists do indeed show preferences for work like their own during peer review. Mahoney (1977) asked reviewers to rate manuscripts with identical procedures, but where conclusions were positive, negative, or neutral, and found a significant bias towards findings supporting the reviewer's own perspectives.[5] Travis and Collins (1991), in an observational study of reviewers of grant applications, found evidence of bias towards one's own "cognitive community"—those sharing the same interests and assumptions. Likewise Lamont et al. (2009), drawing on interviews with panelists from interdisciplinary funding boards, describes preferences towards the practices of panelists' home disciplines. In a more recent and rigorous study of over 600,000 publications, Wang et al. (2017) considered papers with unusual combinations of citations, indicating novelty. Although they found that the most highly cited papers were often novel, novel papers were also systematically published in lower-impact journals, indicating a likely bias against new methods.

We would usually expect new methodologies to spread in a scientific community if researchers can discern a clear advantage of adopting them. For this reason, *competence* to assess methodological quality is critical. By competence we mean the ability of researchers within a particular research tradition to assess the relative quality of methodologies in producing epistemic value. Several studies show low inter-rater reliability among peer reviewers (Cole et al., 1981; Cicchetti, 1991; Mutz et al., 2012; Nicolai

[5]This, of course, shows a preference for a finding, not a method, but still illuminates the sort of self-preferential bias we model here.

et al., 2015) suggesting that competence may be highly variable. While competence will not be the same for all researchers within a discipline, shared traditions, norms, and training practices may create systematic differences of this sort across fields. For example, differences in training in statistical methods or in philosophy of science can leave fields more or less able to differentiate methodological quality, as in the case of MBI.[6] Exacerbating this effect, in some fields the epistemic rewards for adopting new methodologies may not be easy to evaluate, particularly when effect sizes are small and consequences of particular designs take time to be revealed.

3.2. **Interdisciplinarity.** When a scientific community that suffers from strong bias or low competence remains insular, better methods may not spread. However, many scientific communities are linked to each other via shared interests, forming a loose network of communities (Vilhena et al., 2014). Such networks provide a social structure for fruitful interdisciplinary communication. This sort of interdisciplinarity allows different fields or sub-communities to retain their cultural identity and norms, while also receiving and sharing knowledge, ideas, and methods with other communities.[7]

There are different ways this contact occurs. Journals receiving work that cannot be effectively reviewed within their field may seek input from those outside it. Alternatively, a researcher may submit work for evaluation in adjacent fields where their methods are more common. Some fields have norms discouraging such interdisciplinary publication (as with the "top five" journals in economics (Heckman and Moktan, 2020)). If a researcher can be successful by being relatively interdisciplinary though—that is, by publishing in the journals of adjacent fields or by receiving grants centered in adjacent disciplines, or if editors draw on interdisciplinary input in assigning credit within a discipline—it may be possible for better methods to gain traction even in communities marked by strong bias or low competence.

3.3. **Previous Work.** In developing our models, we draw on work from several disciplines. We follow previous authors in assuming that scientists are part of a "credit economy". That is, scientists strive for publications—and the citations, talk invitations, job offers, grant money etc. that ensue—in the same way that normal people strive for wealth or happiness.

Many credit economy models focus on scientists as rational credit seekers. (See, for example, Kitcher (1990); Bright (2017).) Our model, in contrast, falls in line with previous work treating scientists as part of a population where certain behaviors are selected by dint of their success. In particular, as will become clear, we assume that methods which generate credit tend to spread. This could be because prominent role models tend to be imitated, as with Hopkins and Batterham in sports science. It could be due to conscious choices to use methods that generate credit. Or this could stem from the differential success of students whose advisors use credit-producing methods. Previous

[6]For a sardonic examination of low competence among physicists to assess work on theoretical biology, see Shalizi and Tozier (1999).

[7]Of course, with enough time and contact new disciplines may form at the intersections of old ones. But for significant periods of time, structures like departments and journals maintain diversity across academic disciplines, even as interdisciplinary contact occurs.

models have shown that these sorts of selection processes can help explain failures of methodology and discovery in science (Smaldino and McElreath, 2016; O'Connor, 2019; Holman and Bruner, 2017; Stewart and Plotkin, 2020; Tiokhin et al., 2020).

Such models are related to biological models looking at the selection of beneficial behavioral traits. In particular, the multi-group model we present in section 6 bears similarities to a cultural evolutionary model developed by Boyd and Richerson (2002). They consider what happens in a population with different cultural groups who adopt cultural variants via imitation. They assume, as we do here, that imitation tracks payoff success. As they show, beneficial variants can spread between subgroups. We, likewise, are interested in cases where beneficial epistemic practices can spread to other subfields. Unlike those of Boyd and Richerson (2002), however, our models are tuned to details of scientific communities. Also, when it comes to the spread of methods, we focus less on imitation between groups, and more on the possibility that the existence of other disciplines using better scientific methods changes the credit allocations for scientists within a target community.

Akerlof and Michaillat (2018) develop a model to investigate why "false paradigms" persist. That is, why might a scientific community continue to adhere to a set of guiding theories that are sub-optimal from an epistemic perspective? They consider the tenure process, and, like our model, the possibility that scientists are biased towards their own paradigms in tenuring younger faculty. As they show, such a bias can stabilize poor paradigms, especially when tenure judgments are less sensitive to the quality of the work performed. While we make some different modeling assumptions, our findings concerning bias are very similar to theirs, adding robustness to this idea. Unlike Akerlof and Michaillat (2018), though, we investigate the role of outside influences in methodological change. They propose that beneficial paradigmatic change is largely contingent on improvements to the ability of community members to correctly judge the quality of scientists working in different paradigms. This may be the right focus for paradigms that are discipline specific. As we show, however, methods which are used across disciplines can spread through contact between communities.

4. The One-Community Model

We consider a large community where each scientist employs one of two characteristic methods: an "all-right" method, (A), or a "better" method, (B). We assume that the quality of methodology is relevant to the epistemic products of these scientists. The expected epistemic value produced by a researcher using the all-right method is a baseline of $m_A = 1$, while a researcher using the better method produces an expected epistemic value of $m_B = 1 + \delta$. A higher value of δ corresponds to a greater distinction between the quality of the methods. The first question we ask is: under what conditions do these methods yield more or less credit for their users? On the assumption that high-credit methods tend to be imitated, answering this question will allow us to predict whether communities will adopt one method or the other.

Scientists employ their methods to generate results, and are then assigned credit by reviewers in their field. Reviewers are characterized by two field-specific properties:

competence and bias. Competence, ω, is the ability to discern the relative value of another scientist's methods. When $\omega = 1$, reviewers are perfectly accurate in their assessment. When $\omega = 0$, on the other hand, reviewers cannot distinguish the value of a method from the average method used in their field. The competence-based credit rating assigned to scientist i is given by:

$$K_i = \omega m_i + (1 - \omega)\bar{m},$$

where m_i is the quality of the method used by the focal individual and $\bar{m}$ is the mean epistemic quality of the methods used in the population. If we let p be the proportion of the population using better methods, this mean methodological quality is given by

$$\bar{m} = (1 - p)(1) + p(1 + \delta) = 1 + p\delta.$$

Notice that when competence is very low (i.e., $\omega \approx 0$), reviewers rate all papers as having the same quality. In computational versions of the model, detailed below, we can make the model more realistic by adding an error term where reviewer judgments are noisy. In such cases, different papers will receive different quality ratings, but incompetent reviewers will draw all these ratings from the same distribution.

Bias, α, is the extent to which reviewers prefer research that uses similar methods to their own. When $\alpha = 1$, reviewers assign credit entirely based on similarity to their own methods, whereas when $\alpha = 0$, they view methodology solely in terms of its (perceived) objective merit.

The total credit given to research produced by scientist i using method m_i is thus:

$$C_i = (1 - \alpha)K_i + \alpha B_i,$$

where K_i, again, is the contribution based on reviewer competence, and B_i is the contribution based on reviewer bias. B_i will be 1 if the scientist uses the same method as the reviewer, and 0 otherwise.

To review, when α—the weight determining bias versus competence in judgments—is very low and ω—the actual competence of reviewers—is very high, scientists receive a payoff commensurate with the quality of their methods. When ω decreases, reviewers are not biased to a particular method, but are unable to distinguish the underlying quality of the two methods. When α is higher, bias plays a role in review. The credit a researcher can expect to receive becomes dependent on the current distribution of methods in the field, such that more prominently used methods tend to get higher payoffs because more reviewers are familiar with them and prefer them.

5. Payoffs and Invasion in a Single Community

With our model defined by the equations above, we can calculate the expected credit that will be assigned to scientists using either of the two methods. Again, let p be the proportion of scientists using the better method. The expected credit contribution from reviewer bias, B_i, is $1-p$ for an all-right scientist and p for a better scientist. A scientist using method A should thus expect to receive credit of

$$C_A = (1 - \alpha)\left[\omega + (1 - \omega)(1 + p\delta)\right] + \alpha(1 - p),$$

while a scientist using method B should expect to receive credit of

$$C_B = (1-\alpha)\left[\omega(1+\delta) + (1-\omega)(1+p\delta)\right] + \alpha p.$$

If we assume cultural evolutionary dynamics such that scientists who receive more credit will be better positioned to transmit their methods (c.f. McElreath and Boyd, 2007), then method B will increase in frequency whenever $C_B > C_A$. This will occur whenever

$$\delta > \frac{\alpha(1-2p)}{(1-\alpha)\omega}.$$

Method B spreads whenever the epistemic advantage to adopting the better method (δ) is large enough to overcome limitations from the competence and bias of reviewers. Greater competence (ω) lowers this threshold. Greater bias (α) increases it.

The distribution of methods in the community (p) also matters. If more individuals have already adopted method B, bias can work in favor of the better method. In particular, there is often a threshold frequency of users of method B which, if reached for whatever reason, will allow the better method to take over. We can compute this threshold by solving for the minimum proportion of scientists using method B necessary for it to increase in frequency, p^*. This is value of p for which $C_B > C_A$:

$$p^* = \frac{1}{2} - \frac{(1-\alpha)\omega\delta}{2\alpha}$$

If decisions are made entirely based on bias ($\alpha = 1$), then whichever method is more common will spread. As bias goes to zero, the better method is increasingly guaranteed to spread. For intermediate cases, increased competence can move the critical threshold lower, so that a better method held by the minority can still spread even in the presence of bias.

Let us focus on the case in which method A is firmly entrenched in a scientific community, so that almost everyone is using method A. Under what conditions can an objectively better method increase in frequency, or "invade" in the language of evolutionary game theory? We can calculate the criteria for invasion by setting $p \approx 0$. The following inequality shows the minimum epistemic value advantage for method B to spread:

$$\delta > \frac{\alpha}{(1-\alpha)\omega}$$

Figure 1 illustrates how competence and bias both contribute to the spread of better methods into a community that is currently adopting poor ones. This figure shows the minimum competence, ω, needed for method B to spread when rare as a function of bias, for several values of the method's epistemic advantage, δ. The greater extent to which reviewers are biased, the more competence they must possess to accurately distinguish better methods from the average. The smaller the improvement of the method, the less likely better methods are to spread when reviewers are even somewhat biased and imperfectly competent.

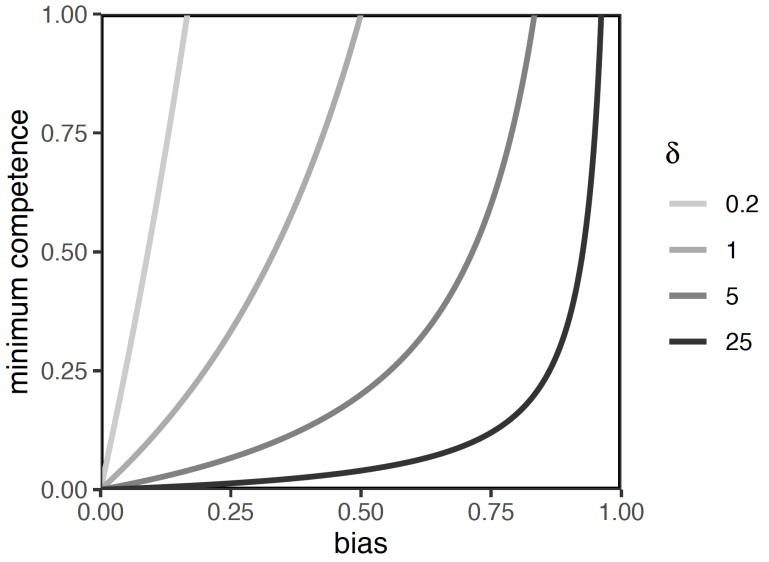

FIGURE 1. Minimum competence, ω, necessary for better methods to invade a population of all-right methods as a function of bias, α. The shape of the curve depends on the epistemic advantage of better methods, δ.

5.1. **Agent-based simulations.** Agent-based simulations allow us to add some realistic noisiness and error to the model. We find that they confirm the findings of the analytic model just presented. (The simulation model described here will also be extended in the subsequent section, when we examine multiple interacting communities.)

In these simulations, we generate a population of $N = 100$ agents. We assume that bias and competence are homogeneous across the group. In each round, agents are assigned credit by a randomly chosen reviewer based on their method, and on the reviewer's bias and competence. Then we use a process similar to the Moran process from evolutionary biology to determine how methods change over time (Moran, 1958). An individual is randomly chosen to adopt a new method. They do so by imitating a group member selected with a probability that increases with that member's credit. In particular, we randomly select five individuals and have them imitate the one with highest credit. (See the Appendix for more detail.) We assume that copiers may occasionally experiment by adopting the method that is not associated with the highest credit, with probability μ. We initialize the population with 95% of agents using method A, and with 5% using method B. We ran simulations long enough that they reached stable states to see whether the population evolved to use the better method over the course of

simulation.[8] A full description of the agent-based model, including all parameters used, is provided in the Appendix.

The top row of Figure 2 shows the prevalence of methods A and B at the end of simulation. As is evident, the analytic predictions (pictured as magenta lines) correspond to the outcomes of these simulations. Bias and competence trade-off in determining whether the better method can invade. And the epistemic difference, δ, shifts the degree of competence necessary to offset bias.

As any academic who has undergone peer review will know, this process is not always a consistent one. For this reason we also ran simulations in which error was introduced into credit assignment. Under this assumption, the credit assigned to scientist i is:

$$C_i = (1 - \alpha)K_i + \alpha B_i + \epsilon,$$

where ϵ is a random draw from a normal distribution with a mean of 0 and a standard deviation of σ. This error term captures two sources of noise in the evaluation process. First, the same method may produce research of higher or lower epistemic value based on the details of the research question and noise inherent in any complex process. Second, a reviewer may be influenced by stochastic factors in addition to general competence and bias, ranging from hunger to prejudice.

We find that noise actually increases the likelihood that better methods will spread for values of bias and competence near the threshold. As is clear in the bottom two rows of Figure 2, there are values for which B takes over beyond the analytic threshold. Noise means that sometimes reviewers will rate B very highly, in spite of their bias against it, or lack of competence to judge it. This stochasticity allows the better method to sometimes achieve high enough representation in the population to take over. Of course, we are focused here on the movement from a population with mostly A to one with B. Because we start at the state where very few agents use the better method, there is an asymmetry in how added stochasticity can influence outcomes. That is, it will tend to push away from the baseline, and towards B. In a world dominated by better methods, on the other hand, enough noise could, in theory, cause those methods to become sufficiently rare so that bias would begin to act in favor of "all-right" methods. Given that even low levels of competence always favor better methods, such a scenario would require very high levels of noise and would even then be very unlikely, but it is not impossible for methods in the model to degrade in this manner.

6. Interactions between communities

We have seen that if bias for existing methods is strong or if competence to assess the quality of new methods is low, low quality methods can persist and better methods cannot spread. However, our conclusion holds for an isolated community, in which community members always evaluate one another. In this section, we consider what happens

[8] When $\mu > 0$, this model will never be fully stable, as agents will occasionally adopt the uncommon method. We refer to states where the model stays at approximately the same distribution of methods, with a very small probability of moving away from it.

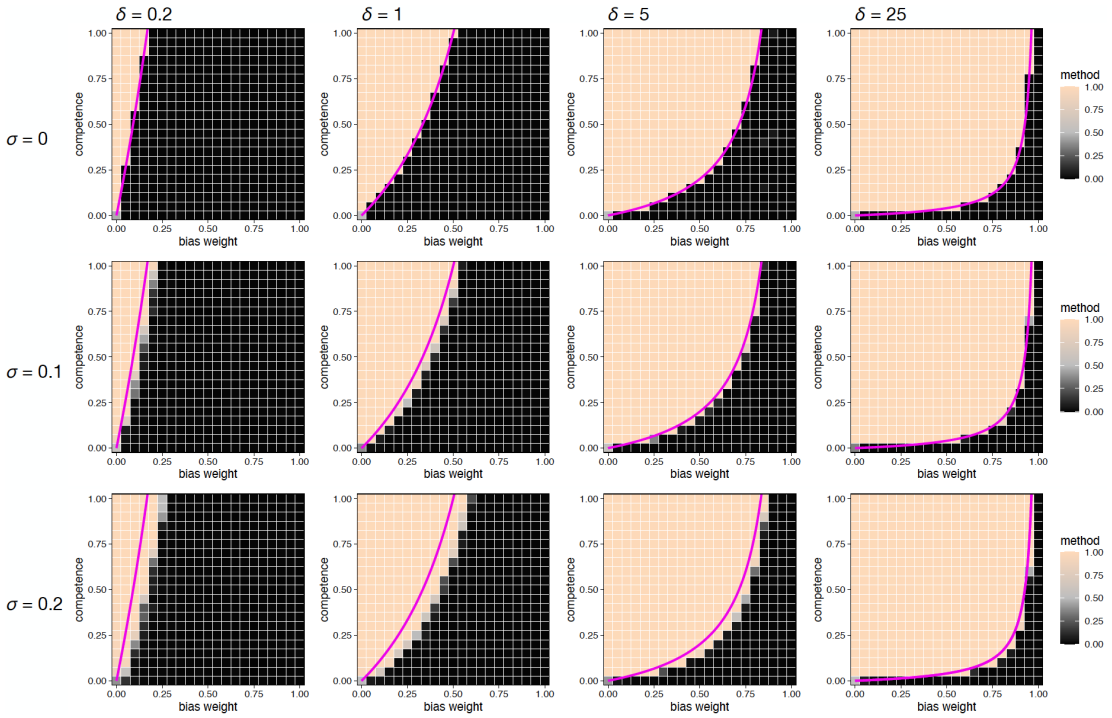

FIGURE 2. Minimum competence, ω, necessary for better methods to invade a population of all right methods as a function of bias, α, for several values of epistemic difference, δ. Colored cells indicate the proportion of agents using method B after 10^5 time steps, averaged across 20 runs. Magenta curves are predictions from the analytical model seen in Figure 1. Increasing the magnitude of noise in credit assignment, σ, enlarges the parameter space for which method B invades.

when scientific communities interact. This model is nearly identical to the one community version, but when a scientist is evaluated for credit, the reviewer is chosen from the other discipline with probability c, which represents the level of interdisciplinarity (see Appendix for more details).

In particular, we are interested in cases where one discipline has adopted the better method, and another the worse one. We initialize community 1 as before, with 95% of agents using method A and only 5% using better methods. We choose parameters such that in the absence of inter-group interaction ($c = 0$), better methods will not spread in community 1. And we initialize community 2 with 95% of agents using method B, and 5% A. We find that across almost all parameter values of this model, a moderate amount of contact between communities leads to the spread of the better method. The rest of this section will elaborate this finding, and point to one edge case where it does

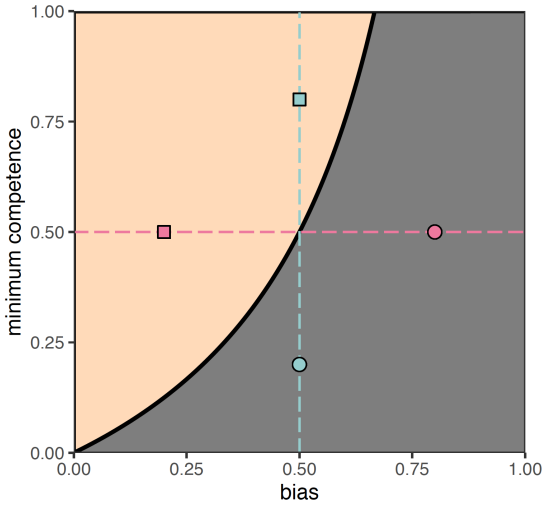

FIGURE 3. The relation between competence and bias for the invasion of better methods within a community, for $\delta = 2$. The gray area denotes a region where better methods cannot invade, the peach-colored area denotes a region where they can. The points show the scenarios comparing all right (circles) and better (squares) communities, with the turquoise points differing on competence ($\omega = \{0.2, 0.8\}$) and the pink points differing on bias ($\alpha = \{0.2, 0.8\}$).

not always hold. For all results presented here, error in credit assignment, $\sigma = 0$, and likelihood of mutation in adoption new methods, $\mu = 0.01$.[9]

We begin by focusing on cases where the method A is common in a community with either (1) low competence or (2) high bias, such that rare users of method B would fail to spread their method if all credit assignment took place within the community. The parameters of bias and competence used to define each community studied in this section are illustrated in Figure 3.

Let us first consider the case in which worse methods are abundant in a community with relatively low competence to assess methodological quality. This could come from better quantitative training or training in the philosophy of science. In particular, we are now interested in the turquoise axis of figure 3. Our two communities have the same bias, $\alpha = .5$, but differ in competence with $\omega = \{.2, .8\}$ for communities 1 and 2 respectively. (Notice this puts community 2 in the regime where the better method will spread absent external factors, and community 1 in the regime where it will not.)

[9]We found that qualitative results were robust across choices of σ, though in some cases it could impact the level of interdisiplinarity necessary for the spread of good methods. Removing mutation so that $\mu = 0$ leads to outcomes where random drift can eliminate the better method in community 1 early on in a simulation. This effect dampens the spread of better methods to community 1, though only by requiring higher levels of interdisciplinary contact.

Simulations show that if a small but sufficient number of credit assignments are made between the communities, better methods can spread within the low competence group. Figure 4A displays these results.[10] The greater the level of out-group credit assignment, the greater the likelihood that the better methods ends up spreading through community 1. In this case, input from community 2 means that those in community 1 using the better method can receive more credit than their colleagues using the all right method, even when the latter method is more common in their community. The result is greater prominence for these individuals, and the spread of their better methods within community 1. This happens for different values of epistemic distinction between the methods, δ.

We next consider the case in which poor methods are abundant in a community with relatively high bias. This could result from a field initially having less accepting norms regarding interdisciplinarity, or from journals standards that look unfavorably upon new or innovative methodologies. The parameters we consider here are represented by the pink axis in Figure 3. Simulation results again demonstrate that increased out-group credit assignment improves the chances that the better method spreads to community 1 (Figure 4B). This may be surprising given that community 1 is strongly biased towards its own methods. But the influence of community 2 overwhelms this bias by assigning credit to higher-quality methods.

Strikingly, the spread of better methods by interdisciplinarity does not actually require that community 2 be particularly competent or unbiased. Instead, if a community adopts good methods even though some accident of history, or through influence from another discipline, they can still pass on these beneficial practices. In Figures 4C and D, we see that even if both communities are incompetent ($\omega_1 = \omega_2 = .2$), or biased ($\alpha = \alpha = .2$) better methods can spread due to interdisciplinary contact.

Figure 5 illustrates that the findings described in this section are robust across parameter values for the two communities. In (A–C) we hold fixed competence with $\omega_1 = \omega_2 = 0.5$ and vary bias for both community 1 and 2. In other words, we further survey parameters displayed by the pink line in figure 3. In each case, the better method spreads when there is enough interdisciplinary contact. In (D–F) we hold bias fixed with $\alpha_1 = \alpha_2 = 0.5$, and vary competence for both communities, thus surveying parameters along the turquoise line in figure 3. In this case as well, enough contact always leads to the spread of better methods. We also ran simulations varying all levels of α and ω flexibly. Although it is difficult to visualize all this data, we find that interdisciplinary contact leads to the spread of better methods in all cases (modulo the edge case described below).

To summarize, interdisciplinary contact, in the form of credit giving between disciplines, can unseat poor methods and replace them with better ones. This finding holds

[10]Unlike the one-population model, we did not run these simulations until they reached stable outcomes (or approximately stable outcomes). This is because with mutation, $\mu = .01$, even after many time-steps it is possible for community 1 to randomly reach a high enough level of better methods for these methods to begin spreading. Therefore, for edge cases, the probability of better methods spreading increased with simulation run time. This fact merely strengthens our claims about the benefits of interdisciplinary contact.

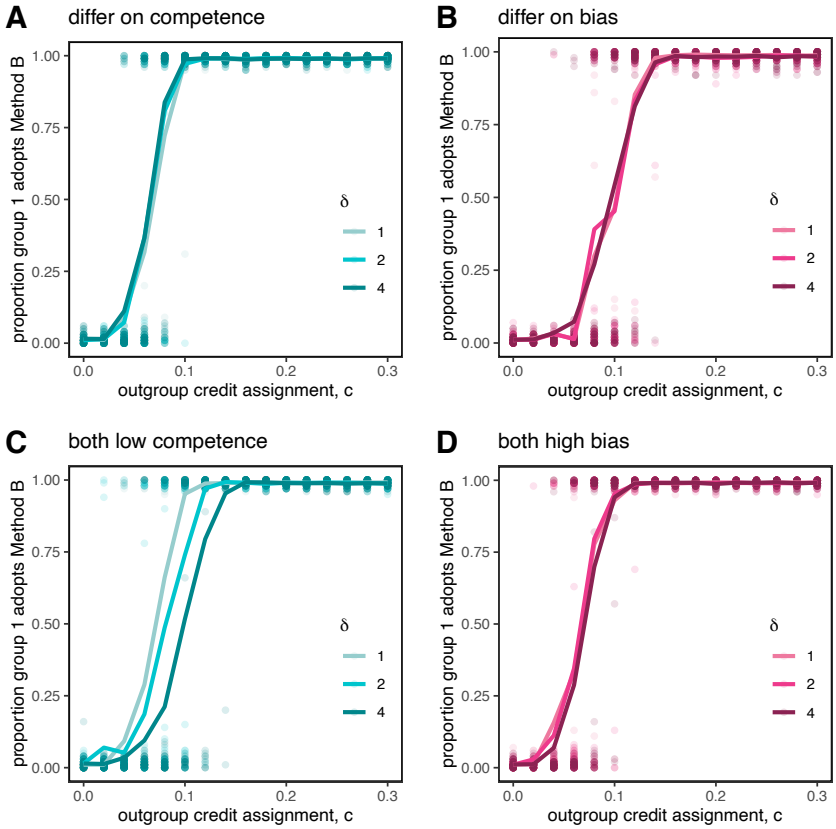

FIGURE 4. Interdisciplinarity allows the spread of better methods into a group that would otherwise not adopt them. Circles represent the proportion of agents in community 1 (the "all right" community) who adopt Method B at $t = 2\times10^5$ time steps, for several values of δ. Lines are the averages across 50 runs for each condition. Top row: Better methods spread from a community with improved norms regarding (A) competence ($\omega_1 = 0.2$, $\omega_2 = 0.8$, $\alpha_1 = \alpha_2 = 0.5$) or (B) bias ($\alpha_1 = 0.8$, $\alpha_2 = 0.2$, $\omega_1 = \omega_2 = 0.5$). Bottom row: Better methods spread from community 2 even when that community is also impaired by (C) low competence ($\omega_1 = \omega_2 = 0.2$, $\alpha_1 = \alpha_2 = 0.5$) or (D) high bias ($\alpha_1 = \alpha_2 = 0.8$, $\omega_1 = \omega_2 = 0.5$).

under a wide range of conditions and is robust to noise (i.e., errors in credit assignment and variability in the quality of findings).

6.1. **Can worse methods ever replace better methods?** We have focused on scenarios in which better methods can spread from a community in which they are common to a community in which they are rare. In none of the scenarios we have examined so far did the reverse ever happen. That is, worse methods never spread into the community

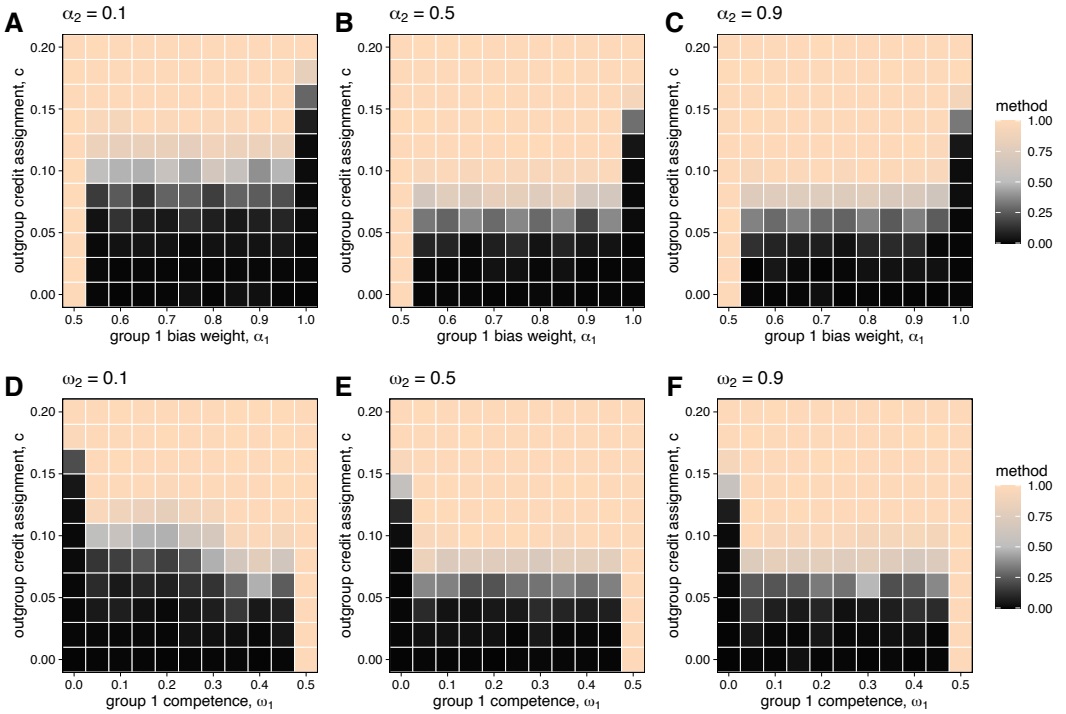

FIGURE 5. The spread of better methods to community 1 is fairly robust to exact values of bias or competence in that community. Colored cells indicate the proportion of agents using method B after 5×10^5 time steps, averaged across 100 runs. We focus on only those parameter regions for which method B will not spread in an isolated community. (A–C) When community 1 is characterized by strong bias, sufficient outgroup credit assignment will nevertheless allow better methods to spread. This is true even if individuals in that community rely entirely on bias in their credit assignments, though more outgroup credit assignment is needed in that case. For these simulations $\omega_1 = \omega_2 = 0.5$. (D–F) Similarly, when community 1 is characterized by low competence, sufficient outgroup credit assignment will allow better methods to spread. If community members have absolutely no competence to evaluate better methods, sufficient outgroup credit assignment can still facilitate the spread of better methods, though more outgroup credit assignment is needed in that case. For these simulations $\alpha_1 = \alpha_2 = 0.5$. For all simulations, $\delta = 2$, $\mu = 0.01$, and $\sigma = 0$.

in which better methods were initially common. Could this ever happen? Our analyses suggest this outcome is highly unlikely, but not impossible. Consider a scenario in which there is widespread lack of competence to compare two different methodologies. That is, scientists in both communities cannot tell if one method is better than another. In

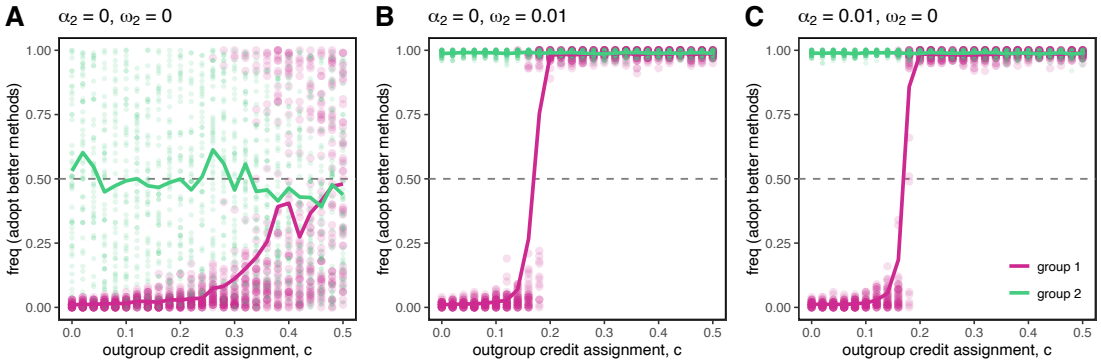

FIGURE 6. Long-run frequency of the better method in each community under the "worst case" scenario. (A) Here both groups lack competence to compare methods ($\omega_1 = \omega_2 = 0$), community 1 is strongly biased toward current methods ($\alpha_1 = 1$), and community 2 lacks any such bias ($\alpha_2 = 0$). Under this scenario, worse methods can infiltrate community 2 and spread by neutral drift. However, even a very small amount of (B) competence or (C) in community 2 halts this effect. For all simulations, $\delta = 2$, $\mu = 0.01$, and $\sigma = 0$. Circles represent the proportion of agents in each community who adopt Method B at 2×10^5 time steps, lines are the means across 50 runs for each condition.

this case, they instead rely on bias. But what if this bias is strongest in communities that have adopted worse methods, and weaker in communities that have adopted better methods?

We analyze this "worst case" scenario by again considering two communities in which the better method is initially rare in community 1 and common in community 2. Both communities here lack any competence to compare methods ($\omega_1 = \omega_2 = 0$), and members of community 1 are strongly biased while members of community 2 are completely open minded ($\alpha_1 = 1$, $\alpha_2 = 0$). We find that under these circumstances, the effect of drift dominates: worse methods permeate into community 2 more readily, although this drift may also facilitate the spread of better methods to community 2 with sufficiently large c. Nevertheless, we view the scenario considered in this section as rare. Our analyses show that even minimal competence or bias in community 2 halts this backslide as Figure 6 shows.

7. CONCLUSION

Strong bias for current methods or low competence to assess new methods can impede scientific progress. As noted earlier in the paper, there are other cases besides the use of MBI in sports science that seem to fit well with the analysis provided here. For instance, the discipline of evolutionary psychology has been widely criticized for sometimes publishing findings that are unsupported and even irresponsible. A central part of these criticisms involves the development of ultimately speculative narratives

about the evolution of the human mind. As critics argue, these narratives are too unconstrained, and thus are unlikely to accurately capture facts about human evolution (Gould, 1991; Lloyd, 1999; Lloyd and Feldman, 2002; Gannon, 2002). When such work is published in journals edited and assessed by evolutionary psychologists, both bias and lack of competence can contribute to the maintenance of these uncritical applications of evolutionary theory. Competing traditions of more rigorous work in the evolutionary human sciences including in human behavioral ecology, gene-culture coevolution theory, and also by some more rigorous evolutionary psychologists (c.f. Gurven, 2018; Amir and McAuliffe, 2020; Barrett, 2020b), point to promising avenues for future reform.

The field of social psychology has recently undergone serious methodological changes. Retrospective studies found that, particularly pre-2011, the field was rife with poorly designed and improperly analyzed data. Many attempts to replicate high profile findings in the field failed (Open Science Collaboration, 2015; Ebersole et al., 2016). Some problematic practices like small sample sizes, lack of credible priors, and use of invalid measurement instruments may have persisted for a long time in this discipline because of the types of factors we identify here.[11]

But not all of the problematic methods in this field and others are well-captured by our models. Consider HARKing (hypothesizing after results are known) and p-hacking (selectively reporting or manipulating data to produce desirable inferential statistics). Both of these are what we might call *implicit methodologies*. They were widespread, not generally recognized as problematic, and were taught to students (and so perpetuated), but were not typically reported in the details of published work. Competence likely played an important role in the persistence of these poor methods, but not because reviewers or editors were incompetent to realize that the methods were poor. Instead, reviewers may have been largely unaware of p-hacking or HARKing in manuscripts they reviewed. Rather, the issue stemmed from the competence of the researchers themselves, who did not have the statistical and/or philosophical training to understand the problems with these practices.[12] And self-preferential biases would likewise have little effect on the acceptance or rejection of papers using implicit methods, because, again, reviewers are unaware of these methods. Subsequent improvements in social psychology have benefited from interdisciplinary contact and feedback, but have nevertheless largely been driven by critiques from within the field. This is to say that the mechanisms we focus on throughout this paper are important, but not obligatory, for explaining the persistence, and eventual improvement, of poor methodology.

Our analysis suggests that interdisciplinary contact should be promoted, as such contact may facilitate the spread of better methods to disciplines where they have not yet been adopted. We have already talked about some keys mechanisms for this contact—the use of competent reviewers from neighboring disciplines, and the ability of academics to receive grants from, or publish in, other disciplines. There are some other mechanisms

[11]There are likely many more cases that fall under the purview of the models here. Documenting the reasons why scientists do or do not adopt new methods is challenging, though. For this reason we do not speculate further about other possible domains of applicability.

[12]Indeed, incentives to publish positive and exciting results may have *selected* for misunderstandings of methodological details (Smaldino and McElreath, 2016).

that might be worth considering, though. The advent of social media platforms has led to increased contact between those in different disciplines, and especially to opportunities for feedback between these groups. This may help to lessen the silo-ing of academic fields. Furthermore, online platforms may improve opportunities for interdisciplinary credit giving. When academics are cited outside their area or invited to high-profile speaking opportunities, publication opportunities, etc. in different areas, their prominence and influence within their own area may subsequently increase. Likewise interdisciplinary conferences, special issues, and institutes may play an important role in the spread of good methods. Such institutions bring academics from different disciplines into contact, increasing chances of credit giving through citations and invitations, and increasing future chances of cross-disciplinary review.

Notice that in our model there is no direct copying across disciplines. We assume that researchers are interested in adopting methods that best suit those in their own field, rather than other fields. But it is certainly the case that researchers *do* imitate across disciplines in many instances. This can be an important source of new methods. In the model of Boyd and Richerson (2002), it is direct copying of other groups that allows the spread of beneficial variants between cultural groups. This hints at another, related way that interdisciplinary contact may improve scientific methodology.

One might conclude from all this that the best structure for scientific communities is then a flat one, without disciplinary boundaries. But our results do not actually support this conclusion. A single unified scientific community would suffer from the problem indicated by our baseline model: the risk that new and improved methods fail to spread. Indeed, many have pointed to the benefits of certain types of diversity in academia. Longino (1990) in particular has advocated for critique across diverse views as important for rooting out poor assumptions and practices in science. Feminist philosophers of science like Longino (1990) and Okruhlik (1994) have lauded personal identity as an important source of the necessary cognitive diversity. Another such source stems from a diversity of educational regimes. However, several forces act to decrease diversity of practice within close-knit communities like academic disciplines, including human tendencies towards conformity, norm following, and practices of indoctrination. Some disciplinary structure may also be important in preserving diversity of methods and assumptions. The aim, though, is to have enough contact between disciplines so that this diversity can prove beneficial to science as a whole.[13]

Given this we might ask: what stands in the way of sufficient interdisciplinarity in science? There are norms against interdisciplinarity in some fields. For instance, some fields consider publication in top insider journals a requirement for promotion. This limits the possibility of interdisciplinary credit-giving. In some cases these norms might arise from in-group favoritism and biases against the out-group. In other cases, they

[13]There is a connection here to research indicating an intermediate amount of contact between groups, rather than full connectivity, is optimal for solving some types of complex problems (Lazer and Friedman, 2007; Derex and Boyd, 2016), and to work suggesting that an intermediate amount of communication is optimal for scientific theory change because it ensures transient diversity of beliefs in science (Zollman, 2010).

might arise out of a desire to preserve the special status of a discipline.[14] Other fields may be silo-ed as a results of an inability to understand or engage with outside disciplines. This implies that improved training may help, such as required graduate courses in methods from nearby fields.

In thinking about these possible reforms, we would like to highlight one questionable assumption that we make. Our models suppose that good methods are no more difficult to employ than bad ones. But in many cases, rigorous methods are difficult to use, and part of the draw of non-rigorous ones is that scientists can publish more quickly and easily (Smaldino and McElreath, 2016; Heesen, 2018). In a regime where there is enormous pressure to publish, this is very attractive. Where poor methods are driven by this sort of process, interdisciplinary contact may be less useful.

Many important scientific advances have come from interdisciplinary connections, particularly when methods or theories find new applications in other fields, or when new amalgamate disciplines (e.g., cognitive science, cultural evolution, network science) emerge from cross-disciplinary consolidation. Our analysis suggests an additional benefit to interdisciplinarity: it may improve the methodological quality of the disciplines involved.

Acknowledgments: Many thanks to Christie Aschwanden, Kristin Sainani, and Aaron Caldwell for communications about methods in sports science. Thanks to Andrew Vigotsky for sharing unpublished work on sports science. For comments on previous drafts of this manuscript, we thank Liam K. Bright, Remco Heesen, Colin Holbrook, Aaron Lukaszewski, Leo Tiokhin, and Pete Richerson. And finally, thanks to the organizers of the Metascience 2019 conference, which inspired this paper.

References

Akerlof, G. A. (2020). Sins of omission and the practice of economics. *Journal of Economic Literature*, 58(2):405–18.

Akerlof, G. A. and Michaillat, P. (2018). Persistence of false paradigms in low-power sciences. *Proceedings of the National Academy of Sciences*, 115(52):13228–13233.

Amir, D. and McAuliffe, K. (2020). Cross-cultural, developmental psychology: Integrating approaches and key insights. *Evolution and Human Behavior*.

Aschwanden, C. (2018). A flawed statistical method was just banned from a major sports science journal.

Aschwanden, C. (2019). Sports science is finally talking about its methodology problems.

Aschwanden, C. and Nguyen, M. (2018). How shoddy statistics found a home in sports research.

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, 41(2):203–226.

Barker, R. J. and Schofield, M. R. (2008). Inference about magnitudes of effects. *International journal of sports physiology and performance*, 3(4):547–557.

Barrett, H. C. (2020a). Deciding what to observe: Thoughts for a post-WEIRD generation. *Evolution and Human Behavior*.

[14] Maintaining such a special status might even be tied to explicit financial incentives, as with economics vis-à-vis the other social sciences.

Barrett, H. C. (2020b). Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends in Cognitive Sciences*.

Batterham, A. M. and Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *International journal of sports physiology and performance*, 1(1):50–57.

Boyd, R. and Richerson, P. J. (2002). Group beneficial norms can spread rapidly in a structured population. *Journal of theoretical biology*, 215(3):287–296.

Bright, L. K. (2017). Decision theoretic model of the productivity gap. *Erkenntnis*, 82(2):421–442.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1):119–135.

Cole, S., Cole, J. R., and Simon, G. A. (1981). Chance and consensus in peer review. *Science*, 214(4523):881–886.

Derex, M. and Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, 113(11):2982–2987.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82.

Gannon, L. (2002). A critique of evolutionary psychology. *Psychology, Evolution & Gender*, 4(2):173–218.

Gould, S. J. (1991). Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues*, 47(3):43–65.

Gurven, M. D. (2018). Broadening horizons: Sample diversity and socioecological theory are essential to the future of psychological science. *Proceedings of the National Academy of Sciences*, 115(45):11420–11427.

Heckman, J. J. and Moktan, S. (2020). Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, 58(2):419–70.

Heesen, R. (2018). Why the reward structure of science makes reproducibility problems inevitable. *The Journal of Philosophy*, 115(12):661–674.

Henrich, J. and Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, 19(4):215–241.

Holman, B. and Bruner, J. (2017). Experimentation by industrial selection. *Philosophy of Science*, 84(5):1008–1019.

Hopkins, W. G. and Batterham, A. M. (2016). Error rates, decisive outcomes and publication bias with several inferential methods. *Sports Medicine*, 46(10):1563–1573.

Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy*, 87(1):5–22.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lamont, M. et al. (2009). *How professors think*. Harvard University Press.

Lazer, D. and Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694.

Lloyd, E. A. (1999). Evolutionary psychology: The burdens of proof. *Biology and Philosophy*, 14(2):211–233.

Lloyd, E. A. and Feldman, M. W. (2002). Evolutionary psychology: A view from evolutionary biology. *Psychological Inquiry*, 13(2):150–156.

Lohse, K., Sainani, K., Taylor, J. A., Butson, M. L., Knight, E., and Vickers, A. (2020). Systematic review of the use of "magnitude-based inference" in sports science and medicine.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

MacDonald, G. Z., Button, D. C., Drinkwater, E. J., and Behm, D. G. (2014). Foam rolling as a recovery tool after an intense bout of physical activity. *Medicine and Science in Sports and Exercise*, 46(1):131–142.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175.

McElreath, R. and Boyd, R. (2007). *Mathematical models of social evolution: A guide for the perplexed*. University of Chicago Press.

Moran, P. A. P. (1958). Random processes in genetics. In *Mathematical proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge University Press.

Mutz, R., Bornmann, L., and Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLOS ONE*, 7(10):e48509.

Nicolai, A. T., Schmal, S., and Schuster, C. L. (2015). Interrater reliability of the peer review process in management journals. In *Incentives and Performance*, pages 107–119. Springer.

O'Connor, C. (2019). The natural selection of conservative science. *Studies in History and Philosophy of Science Part A*, 76:24–29.

Okruhlik, K. (1994). Gender and the biological sciences. *Canadian Journal of Philosophy*, 24(sup1):21–42.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Oreskes, N. (2019). *Why trust science?* Princeton University Press.

Sainani, K. L. (2018). The problem with" magnitude-based inference". *Medicine and science in sports and exercise*, 50(10):2166–2176.

Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., et al. (2020). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*.

Shalizi, C. R. and Tozier, W. A. (1999). A simple model of the evolution of simple models of evolution. *arXiv preprint*, pages adap–org/9910002.

Smaldino, P. E. and Epstein, J. M. (2015). Social conformity despite individual preferences for distinctiveness. *Royal Society Open Science*, 2(3):140437.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society open science*, 3(9):160384.

Smaldino, P. E., Turner, M. A., and Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, 6(7):190194.

Stanford, P. K. (2019). Unconceived alternatives and conservatism in science: the impact of professionalization, peer-review, and big science. *Synthese*, 196(10):3915–3932.

Stewart, A. J. and Plotkin, J. B. (2020). The natural selection of good science. *arXiv preprint arXiv:2003.00928*.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., and Donkin, C. (2019). Is preregistration worthwhile. *Trends in Cognitive Sciences*, 24(2):94–95.

Tiokhin, L., Yan, M., and Morgan, T. (2020). Competition for priority and the cultural evolution of research strategies.

Travis, G. D. L. and Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values*, 16(3):322–341.

Vigotsky, A. D., Halperin, I., Gal, D., and McShan, B. B. (2020). Decreasing dichotomization for the wrong reasons: Interpretation of evidence in sports science and medicine researchers. unpublished MS.

Vilhena, D. A., Foster, J. G., Rosvall, M., West, J. D., Evans, J., and Bergstrom, C. T. (2014). Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1:221.

Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Weatherall, J. O. and O'Connor, C. (2020). Conformity in scientific networks. *Synthese*.

Welsh, A. H. and Knight, E. J. (2015). "magnitude-based inference": a statistical review. *Medicine and science in sports and exercise*, 47(4):874.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17.

Appendix A. Agent-Based Model Description

Consider a population consisting of communities of scientists. We considered cases in which there was always either one or two communities. Each community is made up of $N = 100$ scientists, each of whom keeps track of their group identity and is characterized by a method $m_i \in \{A, B\}$. The true epistemic value of method A is set at 1 without loss of generalization; the true epistemic value of method B is set at $1 + \delta$, so that δ represents the epistemic advantage of method B. Each community is initially characterized by a dominant method, which is used by 95% of its population. The remaining 5% of scientists use the non-dominant method. Each community k is further defined by levels of bias, α_k and competence, ω_k.

The dynamics of the model proceed in discrete time steps, each of which consists of two stages: *Science* and *Evolution*.

A.1. **Science.** In the Science stage, each scientist i performs research using their characteristic method and then is assigned credit for that research by a reviewer j. If there is only one community, the reviewer is naturally drawn from this community. In the case of two communities, the is reviewer is drawn at random from the outgroup community (the community to which scientist i does *not* belong) with probability c, and from i's own community with probability $1 - c$. When $c = 0$, each community is completely insular. When $c = 1$, each community is completely evaluated by the other community, an unrealistic scenario that nevertheless creates conditions for conformity. When c is small but nonzero, we have conditions for interdisciplinarity, in which scientists are occasionally judged by the standards of other communities.

Once a reviewer j is chosen, credit C_i is assigned to scientist i according to the following equation:

$$C_i = (1 - \alpha_j)K_{ij} + \alpha_j B_{ij} + \epsilon,$$

where α_j is the bias of reviewer j's community, and epsilon is an error term that captures sources of noise in the evaluation process, as described in the main text. The value of ϵ is a random drawn from a normal distribution with a mean of zero and a standard deviation of σ. Unless otherwise stated, $\sigma = 0$.

To calculate the competence component, K_{ij}, reviewer j considers the mean epistemic quality used in their community, $\bar{m}_j$, given by

$$\bar{m}_j = 1 + p_j\delta$$

where p_j is the frequency of method B in reviewer j's community. The competence component is then given by

$$K_{ij} = \omega_j m_i + (1 - \omega_j)\bar{m}_j$$

To calculate the bias component, B_{ij}, the reviewer simply considers whether they and the scientist i use the same methods, giving credit only when they match:

$$B_{ij} = \left\{ \begin{array}{ll} 1, & \text{for } m_i = m_j \\ 0, & \text{otherwise} \end{array} \right\}$$

This stage continues until all scientists in all communities have been assigned credit.

A.2. **Evolution.** In this stage we use the logic of cultural evolution similar to that used in Smaldino and McElreath (2016), whereby individuals with more credit are more likely to reproduce their methods. This reflects greater success in attracting and placing grad students and postdocs, as well as in influencing other researchers. In each community, one scientist is chosen at random to "die." This does not have to represent literal death, it could also represent retirement or a change of career. Equivalently, we can think of this as a scientist choosing to learn by imitating a high-prestige scientist. Either way, a spot is now open for a new scientist to join the community. A set of five researchers are chosen at random from the community. From among these, the one with the highest credit score is chosen to reproduce. If multiple scientists from this set have the same high credit score, one is chosen at random from among these. In this manner, credit correlates with evolutionary success. This algorithm has been shown to produce qualitatively similar results to one in which an individual's probability of reproducing is explicitly proportional to their credit score (Smaldino et al., 2019). A new scientist is then created to replace the one that died, inheriting the methodology of the reproducing scientist. However, there is also a small probability of experimentation (or, alternatively, innovation or error). With probability $\mu = 0.01$, the new scientist adopts the method that is *not* used by the reproducing scientist. At the end of this stage, all credit scores are reset to zero.

A.3. **Analysis.** Simulations were run for some length of time-steps. The proportions of agents in each community using either method were recorded. The agent-based model was coded in both Java and NetLogo by both authors to confirm the results. The results shown here represent NetLogo simulations. NetLogo code is available at https://www.comses.net/codebase-release/36619eb5-b522-434f-a2ed-f0cbb952ea5b/. Java code is available upon request.