Reputation Without Practice? A Dynamic Computational Model of the Unintended Consequences of Open Scientist Reputations

Maximilian Linde [1,2], Merle-Marie Pittelkow [1,3],
Nina R. Schwarzbach [4], Don van Ravenzwaaij [1]

[1] Unit of Psychometrics and Statistics, Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

[2] GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

[3] QUEST Center for Responsible Research, Berlin Institute of Health at Charité, Berlin, Germany

[4] Unit of Clinical and Developmental Neuropsychology, Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

Practicing open science can have benefits for the career prospects of individual researchers or labs through higher quality work and increased chances of publication. However, being an outspoken advocate of open science might also indirectly benefit individual scientific careers, in the form of status in a scientific community, decisions for tenure, and eligibility for certain kinds of funding. Therefore, it may be profitable for individual labs to appear to engage in open science practices, without actually putting in the associated effort or doing only the bare minimum. In this article, we explore two types of academic behavior through a dynamic computational model (cf. Smaldino & McElreath, **2016**) of an academic community that rewards open science: (1) practicing open science and/or (2) advocating open science. Crossing these two types of behavior leads to four different kinds of labs and we examine which of them thrive in this academic community. We found that labs that practice and advocate open science dominate in a scientific community that values open science. Implications of the model results are discussed.

Keywords *computational model, cultural evolution, metascience, open science, reform*

Initiatives to improve science often follow times of crisis. For example, the open science (OS) movement originated from a crisis in psychology, referred to as replication crisis, crisis of credibility, confidence, or reproducibility (Spellman et al., **2018**; Pashler & Wagenmakers, **2012**; Baker, **2016**; Ioannidis, **2005**; Open Science Collaboration, **2015**; Simmons et al., **2011**; Wagenmakers et al., **2012**; Fiedler, **2011**). Broadly speaking, the OS movement aims to make the scientific process more transparent, accessible, and reproducible. Practices associated with this movement (OS practices) include preregistration and the use of registered reports to reduce researcher's degrees of freedom (Munafò et al., **2017**; Chambers, **2013**), protocol, data and code sharing to improve reproducibility and replicability (National Academies of Sciences, Engineering, and Medicine, **2019**), and the use of preprints and open access publishing to increase the dissemination and accessibility of research findings (McKiernan et al., **2016**; Mikki, **2017**). Incentives such as "OS badges" rewarding openly sharing data or material (Kidwell et al., **2016**), and preregistration (https://osf.io/tvyxz/) advertise and identify "trustworthy" research (Schneider et al., **2020**). More recently, additional badges for open access publication, open code, open source, and open science grants have been proposed (Guzman-Ramirez et al., **2023**).

There is wide-spread agreement that adopting OS practices has advantages for both science at large as well as the individual researcher (Allen & Mehler, **2019**; McKiernan et al., **2016**; Markowetz, **2015**). Researchers are encouraged to use OS practices to advance their career by increasing their citation

Take-home Message

Labs that practice open science (e.g., preregistration, registered reports, sharing data, materials and codes, and open access publishing) and advocate open science (e.g., through social media) thrive in a scientific community that values open science. At the same time, "quick-and-dirty" science is still prevalent, as evidenced by high false positive and false discovery rates. Based on the specific assumptions of our model, our results suggest that labs that practice and advocate open science are dominating in a scientific community that values open science. These results are encouraging to those who feel practicing open science "is not worth it": in addition to benefits to science at large, our results suggest engaging with open science can benefit individual researchers if open science is sufficiently rewarded.

count, generating media attention, attracting potential collaborators, and getting job and funding opportunities (McKiernan et al., **2016**). Moreover, policy decisions aim to recognize and reward the use of open science (see, e.g., https://www.nwo.nl/en/recognition-and-rewards). However, incentivising OS practices might bring along secondary, unintended problems. As the traditional publish-or-perish culture may have inspired questionable research practices like p-hacking and hypothesizing after the results are known, so too may the elevated status and increased publication chances of practitioners of OS inspire advocating OS without actually engaging in OS practices.

The Present Study

In this article, we explore the benefit of practicing and advocating OS in a scientific community that rewards OS. To this end, we extended a computational model by Smaldino and McElreath (**2016**) who demonstrated that the current incentive structure in science, that rewards many and highly cited publications, could lead to low quality studies. Their results imply that in order to be successful, labs should favor a "quick-and-dirty" approach to conducting studies even though that would lead to a high false positive rate and a high false discovery rate. We extended this computational

model by including four different lab types that are a factorial combination of practicing OS (yes/no) and advocating OS (yes/no). In this work, we (1) examined which lab type(s) dominate(s) in a scientific community that values OS and (2) investigated the dynamics of several characteristics in a scientific culture.

We highlight that this work is exploratory and meant to be a proof of principle. While we ground our operationalizations and the selection of parameter values in the existing literature and our personal experiences as OS researchers, we do not claim that our results fully capture the complexity and the individuality of OS labs. Rather they are a simplification of reality and aim to illustrate how the landscape of science might change under different conditions.

Evolution of Bad Science

The original methods to build the evolutionary model are reported in Smaldino and McElreath (**2016**), and a detailed explanation is provided in Box 1. In short, the model starts with a population of $N = 100$ labs that conduct research, publish papers, and gain rewards based on the number of publications and their associated value. Each lab is characterized by: (1) power W, the ability/probability of a lab to positively identify a true effect; (2) replication rate r, the probability of conducting a replication; (3) effort e, the amount of time a lab spends on conducting a study; and (4) false positive rate α, the probability that a lab incorrectly claims an effect (in statistical testing referred to as the significance level).

Variation in these four characteristics leads to variation in fitness of the labs, which determines which labs "die" (e.g., a principal investigator no longer has any students or funding and as a result decides to leave academia) and which labs "reproduce" (e.g., a prolific PhD-student from a successful lab starts a lab of their own) to create offspring labs. Survival of the labs depends on payoffs that they receive for publishing research projects. At each time step, each lab either initiates a new investigation or not. The new investigation can be either a replication study or not. Results of a new investigation can be negative ($-$) or positive ($+$),

Box 1: Evolution in Smalidino and Mcelreath (2016)

Evolution Characteristics

Evolution takes place over many time steps in the model (i.e., $100,000$ in Figure 3 and $1,000,000$ in Figures 4 and 5 of Smaldino & McElreath, **2016**). At time step 1, the simulations are initialized with lab characteristics $W = 0.8$, $r = 0.01$, and $e = 75$ for each lab.

Probability and Type of Investigation

The probability that a lab launches a new investigation (h) at a given point in time depends on η (the influence of effort on productivity) and e of the corresponding lab:

$$h(e) = 1 - \eta \log_{10}[e]. \tag{1}$$

If the lab initiates a new study at a given time step, it is a replication study with probability r, which varies across labs; it is a novel study with probability $1 - r$.

Probability of Obtaining a Positive Result

If the new study is a novel study, the underlying hypothesis is true with probability b, which is fixed at $b = 0.1$ for all time steps and labs; the underlying hypothesis is false with probability $1 - b$. If the underlying hypothesis is true, the lab observes a positive novel finding with probability W, which varies across labs; if the underlying hypothesis is false, the lab observes a positive novel finding with probability α, which varies across labs.

Probability of Publishing and Payoff

A positive novel finding will be published with probability $C_{N+} = 1$. If the positive novel finding is published, the corresponding lab receives a payoff of $V_{N+} = 1$ that is added to the already accumulated payoff. A negative novel finding will be published with probability $C_{N-} = 0$. If the negative novel finding is published, the corresponding lab receives a payoff of $V_{N-} = 1$ that is added to the already accumulated payoff. Any published novel finding (i.e., both positive and negative) will be added to the literature and is therefore available as a target for replication by other labs.

If the new study is a replication study, a hypothesis is randomly chosen from the literature (i.e., from the collection of studies that were already conducted by other labs). If the underlying hypothesis of the original study is true, the lab observes a positive replication finding with probability W, which varies across labs; if the underlying hypothesis of the original study is false, the lab observes a positive replication finding with probability α, which varies across labs.

A positive replication finding will be published with probability $C_{R+} = 1$. If the positive replication finding is published, the corresponding lab receives a payoff of $V_{R+} = 0.5$ that is added to the already accumulated payoff. Moreover, the lab that originally investigated the hypothesis receives a payoff of $V_{0+} = 0.1$ that is added to the already accumulated payoff. A negative replication finding will be published with probability $C_{R-} = 1$. If the negative replication finding is published, the corresponding lab receives a payoff of $V_{R-} = 0.5$ that is added to the already accumulated payoff. Moreover, the lab that originally investigated the hypothesis receives a payoff of $V_{0-} = -100$ (a penalty) that is added to the already accumulated payoff.

Evolution Dynamics

At each time step, the mean of W, α, and e across labs, and the false discovery rate (FDR) are calculated. FDR corresponds to the proportion of false positive findings among all positive findings across labs at a given time step. However, data is only collected (written to a file) every $2,000$ time steps.

After every time step, an evolution step takes place in which one lab ``dies'' and one lab ``is born''. To determine the dying lab, $d = 10$ labs are randomly selected, of which the lab with the highest number of active time steps dies. If multiple labs tie, one is chosen at random. To determine the lab that procreates, $d = 10$ labs are randomly selected, of which the lab with the highest payoff reproduces. If multiple labs tie, one is chosen at random.

The offspring lab inherits the characteristics from the reproducing lab. However, the inherited characteristics are allowed to mutate. All characteristics (r, e, W) mutate with a probability of $\mu_r = \mu_e = \mu_W = 0.01$. If characteristics do mutate, the new value is $\mathcal{N}(x, y)$, where x corresponds to the old value of the characteristic and y to either 0.01 for r and W or 1 for e. If this mutation process exceeds a boundary of the allowed range ($[0, 1]$ for r and W; $[1, 100]$ for e), the corresponding boundary is used as the new characteristic value.

Lastly, if labs publish novel studies, they are added to the literature. The size of the literature is limited to $1,000,000$ hypotheses. If the number of hypotheses in the literature exceeds $1,000,000$, the oldest hypotheses are removed until the number of hypotheses in the literature is $1,000,000$ again.

which determines their probability to be published C. The payoff for a published result V depends on whether it is a novel (N) or a replication (R) study and whether its outcome is negative ($-$) or positive ($+$).

▌ Extension

We extended the model of Smaldino and McElreath (**2016**) by differentiating between labs that do or do not practice OS and between labs that do or do not advocate OS, yielding four types of labs (see Table 1).

The four lab types were initially represented in equal proportions (i.e., at time step 1). When a lab reproduced, the offspring lab automatically inherited the lab type from the reproducing lab. To have enough labs of each category, we increased the number of labs from $N = 100$ to $N = 400$.

We made the following assumptions about the impact of practicing OS on the survival of labs:

1. **Practicing OS leads to higher workload (e).**

   The practice of OS requires more work and time compared to closed science (Hostler, **2023**). New skills and knowledge need to be acquired and the research process involves additional steps, such as pre-registration, data and code cleaning, and additional administration (e.g., drafting openness agreements Hostler, **2023**). Indeed, practicing OS is associated with an increase in workload, work-related stress, and longer time to completion of a research project (Sarafoglou et al., **2022**; Toth et al., **2021**). Seen as "increasing effort decreases the productivity of a lab, because it takes longer to perform rigorous research" (Smaldino & McElreath, **2016**, p. 6), we reasoned that labs that practice OS should have a higher e than labs that do not practice OS. If, for instance, a traditional study were to take 400 work hours to be completed, we assumed that practicing OS would add 20 hours. This translates into a 5% increase in effort for OS studies (i.e., $(400 + 20)/400 = 1.05$). We believe this to be a conservative estimate of the increase in e.

2. **Practicing OS increases the probability of publishing negative novel findings (C_{N-}).**

Table 1 The four different types of labs

		Practicing OS	
		yes	no
Advocating OS	yes	Practice; advocate	Practice; not advocate
	no	Not practice; advocate	Not practice; not advocate

The proportion of published findings with statistically non-significant results is higher for registered reports (60.5%; Allen & Mehler, **2019**) or preregistered studies (52%; Toth et al., **2021**) compared to traditional research, with estimates ranging from 0% to 20% (e.g., Allen & Mehler, **2019**; Fanelli, **2012**). In what follows, we take the liberal estimates: 60% of non-significant OS studies get published and 20% of non-significant traditional studies get published. Assuming the same absolute number of published significant studies in both fields, this means that for every eight statistically significant traditional studies two statistically non-significant traditional studies get published (80% vs 20%); and for every eight statistically significant OS studies twelve statistically non-significant OS studies get published (40% vs 60%). Taking the ratio of statistically non-significant OS studies to statistically non-significant traditional studies, we find that for every non-significant traditional study that gets published, six non-significant OS studies get published. In the original model, the probability of publishing a novel non-significant finding was $C_{N-} = 0$. In our extension, we increased this to $C_{N-} = 0.05$ for traditional studies. To incorporate the six-to-one ratio of non-significant studies between traditional and OS, we set $C_{N-} = 0.3$ for OS studies.

3. **Practicing OS leads to papers that are rewarded more (V_{N+}, V_{N-}, V_{R+}, V_{R-}).**
Citation advantages have been observed for several OS practices. In a systematic review, Langham-Putrow et al. (**2021**) identified 64 studies that claim a citation advantage, 37 studies that do not claim a citation advantage, 32 studies that claim a citation advantage in some subfields, and one inconclusive study (see Table 1 in their article). We used these numbers to approximate a value for

the citation advantage. We only considered the 64 studies that claim an effect and the 37 studies that do not claim an effect. We assumed that the number of citations for OS papers and non-OS papers come from two Normal distributions. Let X be a random variable of $NOS \sim \mathcal{N}(1, 0.1)$ and let Y be a random variable of $OS \sim \mathcal{N}(c, 0.1)$. Through numerical optimization, we found c such that $P(Y < X) = 64/(37 + 64)$. We found an optimal value of $c = 1.0483$, which led to a citation advantage of:

$$\frac{\frac{c - \mu_{NOS}}{\sigma} + 1}{\mu_{NOS}} = \frac{\frac{1.0483 - 1}{0.1} + 1}{1} = 1.483 \tag{2}$$

We found a 48.3% (i.e., 1.483) citation advantage and used this value in our extended model. Note that this is a non-parametric approach to converting the 64 studies that claim an advantage and 37 studies that claim no advantage into a numeric value. Note also that in this approach, the 37 no-advantage studies are operationalized as OS studies being disadvantaged in terms of citation rate.

We made the following assumptions about the impact of advocating OS on the survival of labs:

1. **Advocating OS leads to spending more time advocating (e.g., on Twitter) and less time doing research (η).**
Advocating OS might lead to less available time for doing research because some proportion of the work time is spent on profiling oneself (e.g., posting on Twitter). Therefore, labs that advocate OS had a higher η than labs that do not. We assumed that labs that advocate OS spend two hours of their work time per week (40 hours) on social media.

Table 2 Parameters for the four lab types.

Par.	Value			
	Practice; advocate	Practice; not advocate	Not practice; advocate	Not practice; not advocate
η	$\{0.205, \mathbf{0.210}, 0.215\}$	0.200	$\{0.205, \mathbf{0.210}, 0.215\}$	0.200
e	$\{77, \mathbf{79}, 81\}$	$\{77, \mathbf{79}, 81\}$	75	75
C_{N-}	$\{0.175, \mathbf{0.300}, 0.425\}$	$\{0.175, \mathbf{0.300}, 0.425\}$	0.050	0.050
V_{N+}, V_{N-}	$\{1.543, 1.841, 2.142, \mathbf{2.199}, 2.558, 2.976\}$	$\{1.242, \mathbf{1.483}, 1.725\}$	$\{1.242, \mathbf{1.483}, 1.725\}$	1.000
V_{R+}, V_{R-}	$\{0.771, 0.921, 1.071, \mathbf{1.100}, 1.279, 1.488\}$	$\{0.621, \mathbf{0.742}, 0.863\}$	$\{0.621, \mathbf{0.742}, 0.863\}$	0.500

Parameter values for main analyses are shown in bold font; parameter values for sensitivity analyses are shown in regular font. η is the influence of effort on productivity; e is effort; C_{N-} is the probability of publishing a negative novel finding; V_{N+} is the payoff for publishing a positive novel finding; V_{N-} is the payoff for publishing a negative novel finding; V_{R+} is the payoff for publishing a positive replication; and V_{R-} is the payoff for publishing a negative replication. Par. = Parameter.

It does not matter to the model how much additional time they spend on social media in their free time. We therefore believe that η increases by 5% when advocating OS (i.e., $40/(40-2) = 1.05$).

2. **Advocating OS leads to papers that are rewarded more (V_{N+}, V_{N-}, V_{R+}, V_{R-}).** We assume that publications from labs that advocate OS are rewarded more because they might be read and cited more often. For example, papers that are shared on Twitter (as done by many OS advocates) have a citation advantage over papers that are not shared (Ladeiras-Lopes et al., **2020**; Luc et al., **2021**). We did not find studies specifically focusing on the citation advantage for sharing OS papers on Twitter or other platforms, so we decided to use a heuristic of equating the payoff advantage for labs that advocate OS with the payoff advantage for labs that practice OS (see above; i.e., 48.3%).

The parameter values of our model extension are summarized in Table 2. To make sure our results are robust and not contingent on specific choices for parameter values, we included two additional parameter values for each parameter and ran the factorial combination of each of these as sensitivity analyses (see Table 2). The parameter values of the sensitivity analyses are always 50% and 150% of the

difference between the main parameter values and the reference parameter values.

As a primary result, we collected data on which lab type(s) survive over time and which die out in a world where everyone "plays the game". Specifically, we investigated in what proportions the lab types are present over time: Which lab type(s) is/are most successful within the academic community? As a secondary result, we collected similar data as Smaldino and McElreath (**2016**) about the mean e, mean r, mean α, FDR, and mean W across all lab types. For our simulations, we kept e fixed within lab type (see Table 2).

To reduce the computation time of the simulations, we used a maximum literature size of $100,000$ instead of $1,000,000$. Moreover, we simulated over $1,000,000$ time steps. We sampled every time step for iterations between 1 and $1,000$ iterations, every 10th time step for iterations between $1,000$ and $10,000$, every 100th time step for iterations between $10,000$ and $100,000$, and every $1,000$th time step for iterations between $100,000$ and $1,000,000$.

Further Exploration

We ran an additional set of simulations that incorporated a few more changes to the computational model. First, we reasoned that the payoff for negative novel studies (i.e., V_{N-}) should probably not be as high as the payoff for positive novel studies (i.e., V_{N+}). In an attempt to

Version 1 - Cumulative proportions of lab types
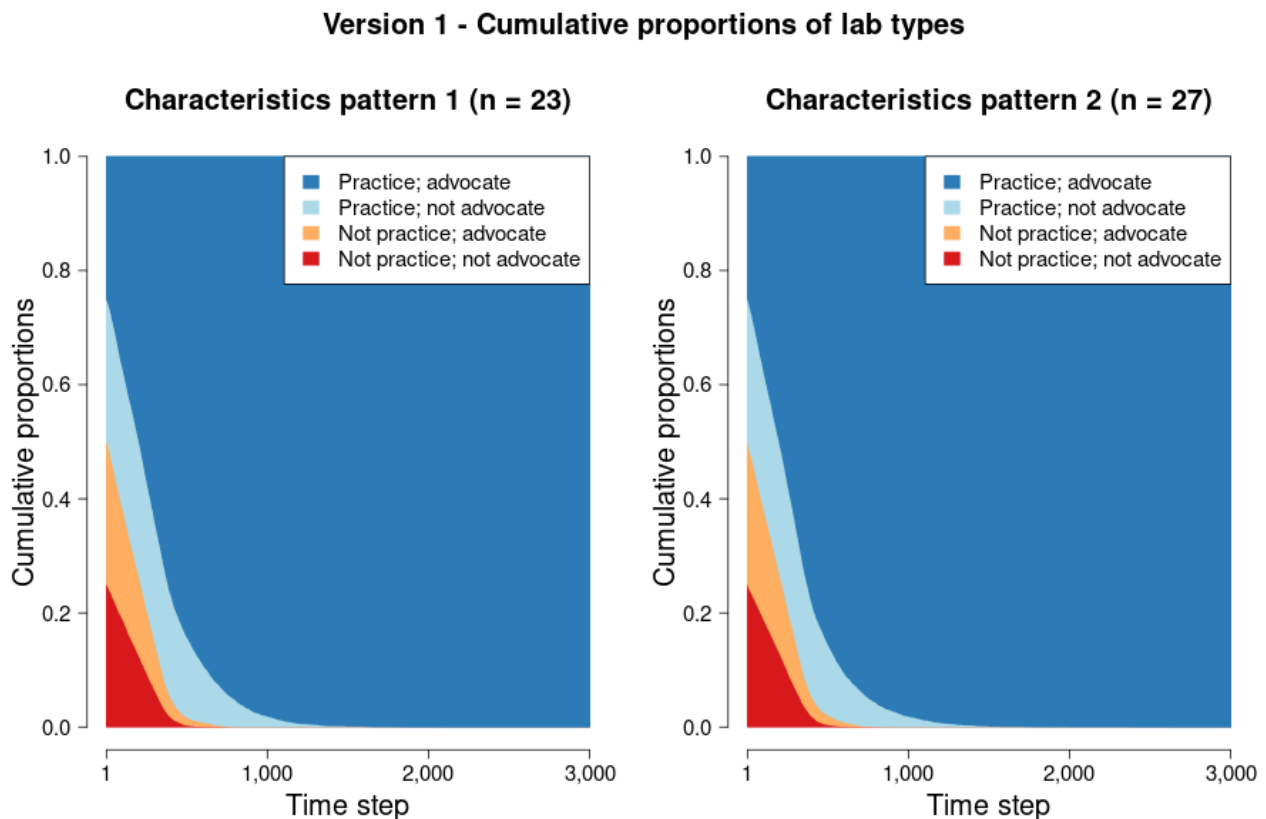
Figure 1   Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the first set of simulations. The two panels represent simulation runs with two qualitatively different patterns of characteristics.

get a more informed estimate, we quantified publication advantage using articles published in the New England Journal of Medicine in 2015. Data was previously extracted by Hoekstra et al. (**2018**). In determining whether a study was considered positive or negative, we focused on statistical inference for the primary outcome. We excluded case studies, descriptive studies, non-inferiority trials, and single-arm studies. Next, we counted all citations (as counted through Google Scholar on date December 9, 2022) for the 120 positive results papers and the 42 null result papers. In the past seven years, null result papers were cited a median of 601.5 times and positive trials were cited a median of 826 times. The ratio of these two medians is 0.728 (i.e., positive novel results have a citation advantage of 1-to-0.728). Therefore,

we set $V_{N+} = 1$ and $V_{N-} = 0.728$ for labs that do not practice and do not advocate OS.

Second, the results of the previous simulations suggest that the characteristics of the scientific community have not yet reached a steady state after $1,000,000$ time steps (see Figure 2). To investigate this further, we decided to increase the mutation probabilities for r and W from $\mu_r = \mu_W = 0.01$ to $\mu_r = \mu_W = 0.1$, so that lab characteristics change more quickly.

❙ Results

Looking at the development of the lab characteristics in the 50 simulation runs separately, we noticed that individual simulations resulted in one of two qualitatively different patterns of characteristics (i.e., FDR, W, α, and r), which are described below. To acknowledge this dichotomy, instead of averaging all 50 simulation

Version 1 - Community characteristics
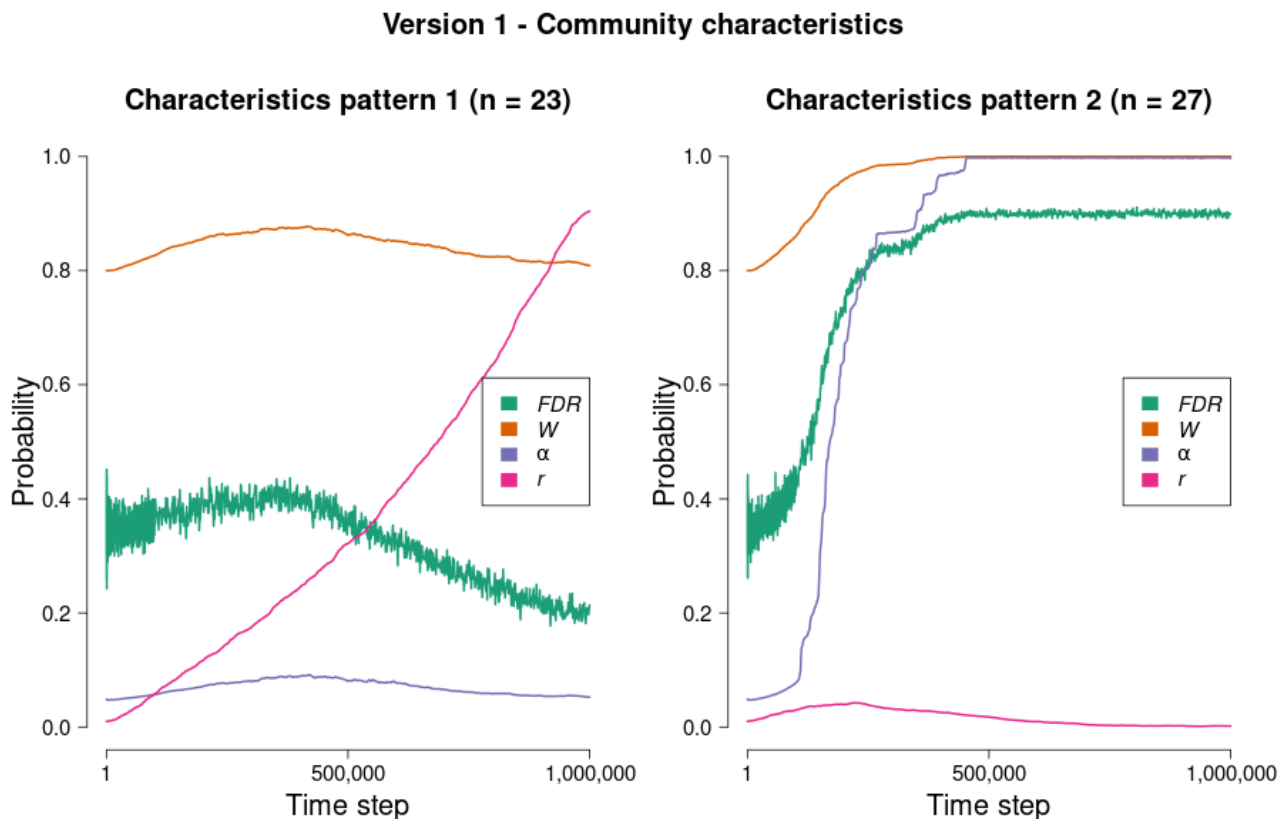
Figure 2   Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over labs and simulation runs over all $1,000,000$ time steps of the first set of simulations. The two panels differentiate between simulation runs with two qualitatively different patterns of characteristics.

runs, we split them according to the pattern of characteristics observed and averaged them separately. This applies to both the proportions of lab types as well as the community characteristics. Accordingly, for our results, we always provide one plot for simulations that exhibited characteristics pattern 1 and one plot for simulations that exhibited characteristics pattern 2.

Who Dominates Science?

Figure 1 shows the proportions of the four lab types over the first $3,000$ time steps. The two panels differentiate between simulation runs that displayed two qualitatively different patterns of characteristics, explained in the next section (see also Figure 2). Labs that practice and advocate OS reach a proportion of 1 very quickly; at the same time, the other three lab types vanish. Of those, labs that do not practice OS and do not advocate OS disappear most quickly, followed by labs that do not practice OS but advocate OS and labs that practice OS but do not advocate OS.

The observed behavior indicates that practicing OS is more important than advocating OS, but that doing both is most advantageous. This advantage of practicing over advocating holds across the entire range of parameter values we investigated (see Figures 5, 6, and 7 in Appendix A.1). The explanation for this is that there is an additional advantage for labs that practice OS, which is the higher probability of publishing a negative novel finding (i.e., C_{N-}, 0.3 versus 0.05). Furthermore, the same behavior is observed for various values of V_{0-} ($-5, -25, -50, -100$; see Figure 17 in Ap-

Version 2 - Cumulative proportions of lab types
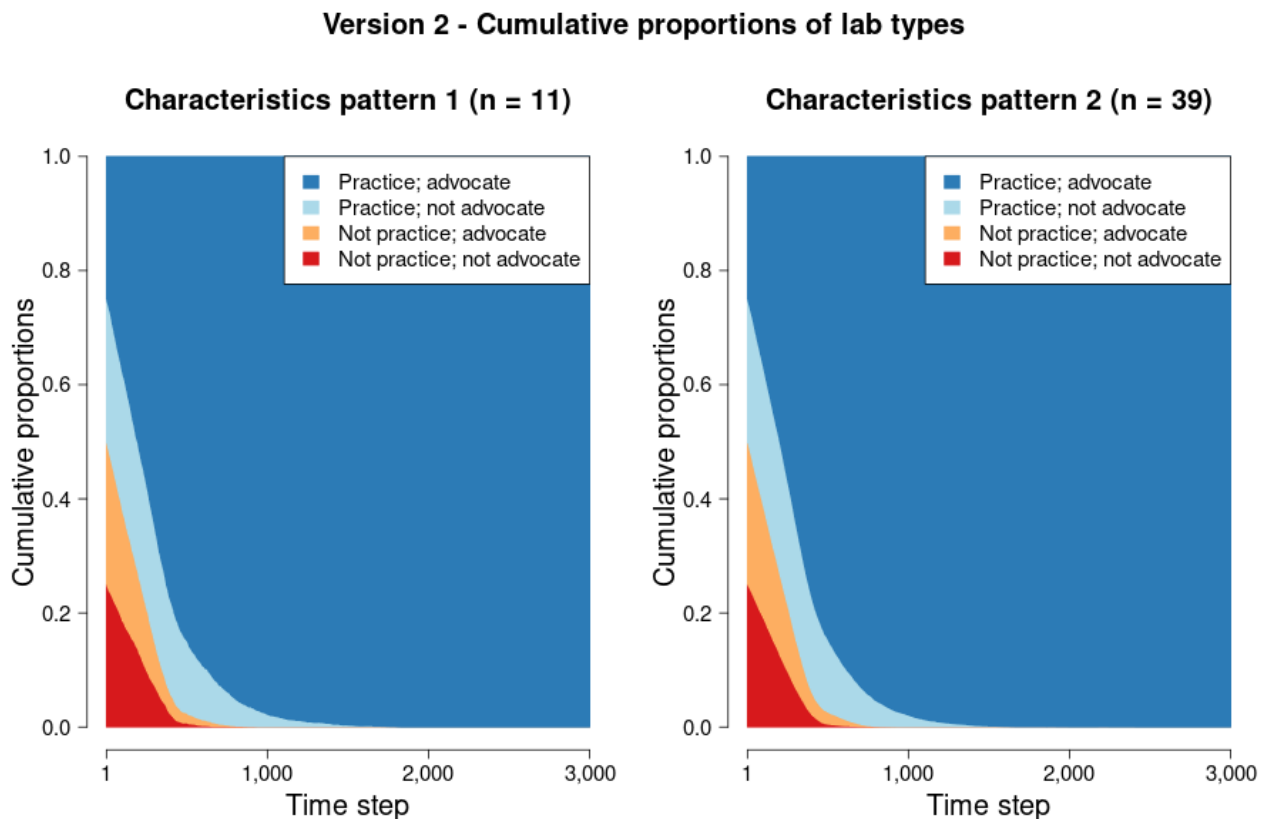
Figure 3 Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the first set of simulations. The two panels represent simulation runs with two qualitatively different patterns of characteristics.

pendix A.2). In sum, in our evolutionary model the additional cost in terms of increased effort (practicing) and a reduction of work time spent on the actual research (advocating) is easily outweighed by the increased payoff when the work gets published.

Characteristics of the Scientific Community

Figure 2 shows the development of characteristics across lab types over the whole range of $1,000,000$ time steps. As all lab types except for the "practice; advocate" lab type ceased to exist within around $3,000$ time steps, the development of characteristics shown in Figure 2 almost exclusively reflects the "practice; advocate" lab type. Here, simulations can be differentiated by two qualitatively distinct patterns of characteristics. In the left panel, it can be seen that W increases slightly and then decreases

very slowly to the initial value. Similarly, α remains fairly constant. FDR rises a bit and then declines over time. Lastly, r increases strongly to almost 1 in an almost linear fashion. An explanation for this is that r increases to a critical value, at which point the values of W and α do not matter (enough) anymore. Recall that for a replication, the type of result does not matter in terms of payoff. As such, r grows to 1. In this variant, mutations are such that the certain payoff of replications was higher than the variable payoff of novel results, with a relatively high occurrence of null results. The reader may notice that W does not grow quite so strongly in the first time steps, giving r enough time to gain momentum.

The right panel displays entirely different characteristics. Here, both W and α increase rapidly to 1 and remain constant. FDR also
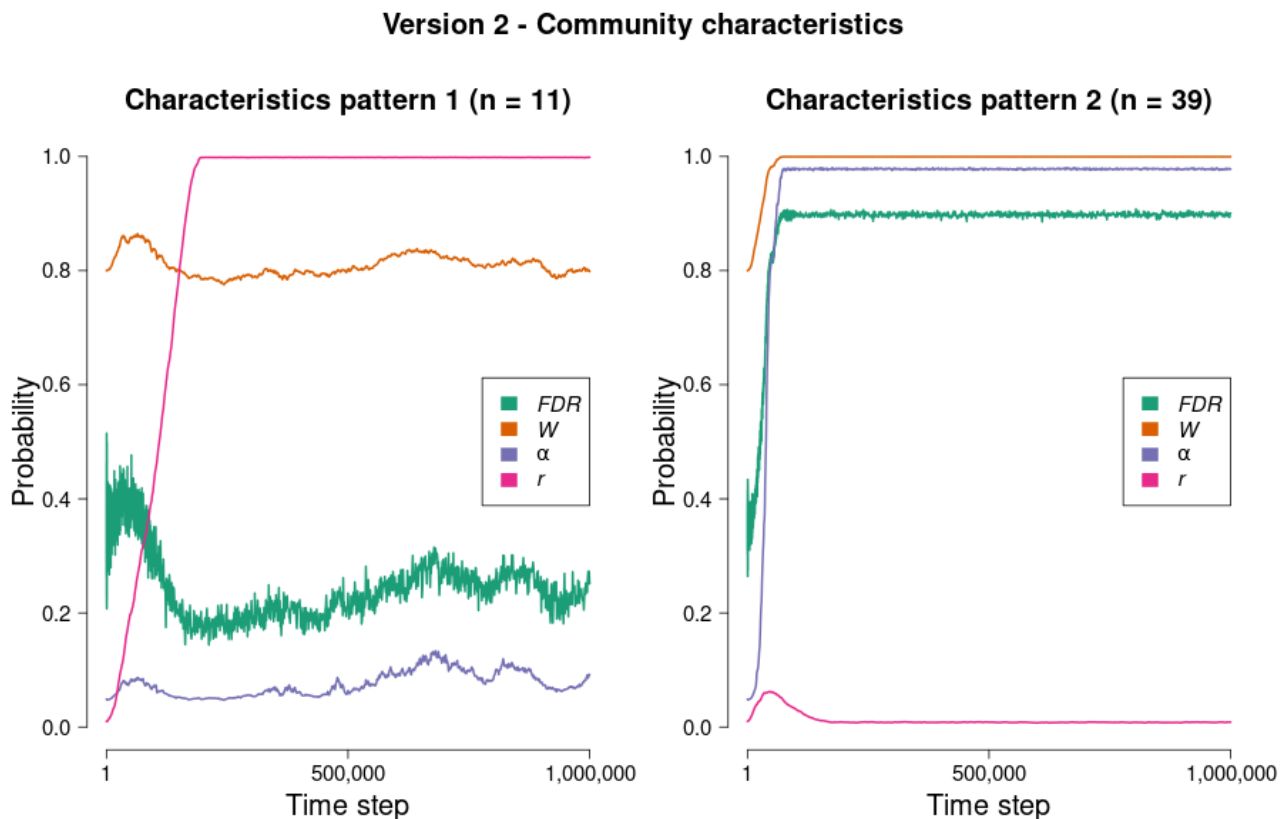
Figure 4 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over labs and simulation runs over all $1,000,000$ time steps of the second set of simulations. The two panels differentiate between simulation runs with two qualitatively different patterns of characteristics.

increases strongly but reaches a plateau at around 0.85. In contrast, r remains very low at almost 0. The explanation is straightforward: if W and α are 1, all results are positive results by default, meaning that every lab should conduct novel studies over replications. This is the case because the payoff is double, and there is no drawback. Note that this second pattern of results was not obtained in Kohrt et al. (**2022**) as they fix power in their simulations.

Further Exploration

We further explored the proportions of lab types and the characteristics of the scientific community with some slight parameter modifications: Changing μ_W and μ_r from 0.01 to 0.1 and changing V_{N-} from 1 to 0.728. Figure 3 shows the proportions of lab types over time.

The behavior is very similar to the previous simulations in Figure 1, such that labs that practice and advocate OS dominate very quickly while the other lab types die out. Once again, these results are robust to different choices of parameter values (see Figures 11, 12, and 13 in Appendix A.1).

Figure 4 shows the characteristics of the scientific community. As in the first set of simulations (see Figure 2), the characteristics reflect those of the labs that practice and advocate OS through most of the time. As in the first set of simulations, we observed two qualitatively different patterns of characteristics of the labs. The explanation for this is the same as for the first set of simulations (see previous section).

Discussion

Science is not just about the academic work– it is ultimately a joint enterprise by people. People who depend on their academic position for their livelihood. As such, doing well, or at least doing better than others, on whatever metric is used to evaluate one's success becomes important to people. Smaldino and McElreath (**2016**) demonstrate that if all people do is "play the game", the scientific work their field produces over time gradually degenerates to low-effort, quick-and-dirty work with a high proportion of false positives.

The OS movement should restrict the feasibility of some of the quick-and-dirty strategies to which researchers might, inadvertently, fall prey to. For instance, preregistering one's work makes it very difficult to employ p-hacking or to hypothesize after the results are known. That said, practicing OS brings with it its own set of success indicators, such as prestige in the field, exclusive funding opportunities, and increased visibility of the work; it may therefore well be optimal to continue to "play the game", just with a slightly adjusted set of rules. In other words, exploiting the incentive structure for OS practices might lead to receiving the same advantages as actually practicing OS. In this study, we explored how different types of academics would thrive in a scientific system that values OS practices. Namely, we compared labs that practice or do not practice OS and labs that advocate or do not advocate OS.

In an incentive structure that values OS practices, practicing OS while also advocating OS is most advantageous. Our simulation results suggest that labs that follow OS practices and engage as "OS advocates" on Twitter or related social media platforms have a survival advantage. The cost associated with both practicing and preaching in terms of a slower "rate of completion" of research projects gets outweighed by the increase in payoff for publications. Within the simulation, only labs that both practiced and advocated OS persisted, all other types quickly vanished from the scientific landscape (i.e., they were less successful in terms of attracting attention and gathering citations).

In our model, advocating OS did not have the same advantages as practicing OS. This was true even in the condition where advocat-

ing was worth more than practicing in terms of publication payoff (73% versus 48%, respectively). The likely reason for this lies in the probability of being able to publish negative or null results: in our parameterization, this probability was six times higher for OS work. Many journals that align with OS principles vouch to judge the publishability of a work based on the soundness of the research question, the methodology, and the study protocol. More traditional journals, in addition to these criteria, tend to lean heavily on the (statistical) significance of the results.

Practicing and incentivizing OS practices did not eliminate quick-and-dirty science in our simulation. While we observed slightly lower values for the false discovery rate and false positive rate compared to Smaldino and McElreath (**2016**) and the replication (Kohrt et al., **2022**), the values were still considerably high (0.7 compared to > 0.8; but see Appendix A.1). Without changing the incentive structure that is currently valuing quantity over quality, OS cannot prevent the rise of quick-and-dirty science. As Simine Vazire once put it (Vazire, **2020**), OS does not act as quality control itself but enables quality control.

Limitations

Although in our model, practicing and advocating OS practices translated to career advantages, some factors cast doubt on the extent to which these findings translate to the real world. Our operationalization of practicing OS involved (slightly) more work per project and a substantial increase in pay-off per publication. Although our parameter settings were grounded to some extent on previous literature, the exact size will be no more than a rough estimate. Perhaps an increase in workload of, say, 50% would be more realistic than an increase of 5%, making the practicing of OS far less attractive for purposes of furthering one's career.

For our operationalization of advocating OS, we assumed that scientists spend two hours of their working weeks on their social media of choice in lieu of working to build and maintain their OS profile. Perhaps two hours is unrealistic and ten hours gets closer to the truth. Or perhaps there is no difference at all in hours spent working between active social media scientists and those that are not active on social

media: time on social media could be spent entirely during free time.

Similarly, it can be argued that the payoff advantage for labs that practice and/or advocate OS is too high and that it is unrealistic that the payoff advantage remains constant throughout evolution. It is possible that it is more realistic for the payoff advantage to diminish over time as an increasing amount of conducted studies are OS studies.

An additional limitation is the omission of consideration regarding the potential consequences of disclosing specific research data. Opening up access to such data may facilitate the identification of inaccuracies and provide a basis for heightened scrutiny from the broader research community (Allen & Mehler, **2019**). Theoretically, this increased transparency could adversely affect the sustainability of a research laboratory, particularly if (unintentional) errors are unveiled and subject to public discourse.

Another, more general, limitation of our setup was the generic classification of scientists as OS practitioners versus non-practitioners and advocators versus non-advocators. In the real world, these categories (and the payoffs they entail) will rarely be so black-and-white. In addition, there will be individual differences in academic success that are tangential to the four categories specified here due to field of interest, background, social network, and even luck. In our simulation, these natural sources of variation were all completely equated. As such, the results of this study should be thought of more as a proof of concept in a drastically simplified representation of what in reality is a very complicated academic ecosphere.

Lastly, our modeling approach to investigate what lab types survive in a community that values open science is only one of many. Different approaches may shed light on the conditions under which labs with different strategies flourish. For instance, using a game theory approach would model the individual labs as rational agents with (potentially) different strategies (e.g., affinity to OS, payoff, etc.). In such an approach, the individual labs would play the "science game", which would allow the computation of an equilibrium distribution of different types of labs.

Conclusion

In our simulation, labs that practice and advocate OS thrive in a scientific community that values OS. At the same time, "quick-and-dirty" science is still prevalent, as evident by high false positive and false discovery rates. These results are encouraging to those who feel practicing open science "is not worth it": in addition to benefits to science at large, our results suggest engaging with OS benefits the individual researcher as well.

▎ Acknowledgements

▎ Protocol, Code, and Data Availability

A transparency documentation of our research process, the code for the simulations, and the data of the simulations can be found online at https://osf.io/h5tfv/.

▎ References

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), 1–14. https://doi.org/10.1371/journal.pbio.3000246 (see pp. 20, 23, 30).

Baker, M. (2016). Dutch agency launches first grants programme dedicated to replication. https://doi.org/10.1038/nature.2016.20287 (see p. 20).

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016 (see p. 20).

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7 (see p. 23).

Fiedler, K. (2011). Voodoo correlations are everywhere - not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171. https://doi.org/10.1177/1745691611400237 (see p. 20).

Guzman-Ramirez, L., Schettino, A., Sweeney, J., & Sunami, N. (2023). Badges to reward open & responsible research practices [Publisher: Zenodo]. https://doi.org/10.5281/zenodo.8278785 (see p. 20).

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PLoS ONE*, *13*(4), e0195474. https://doi.org/10.1371/journal.pone.0195474 (see p. 25).

Hostler, T. J. (2023). The invisible workload of open research [Publisher: JOTE Publishers]. *Journal of Trial & Error*. https://doi.org/10.36850/mr5 (see p. 22).

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124 (see p. 20).

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456 (see p. 20).

Kohrt, F., Smaldino, P. E., McElreath, R., & Schönbrodt, F. D. (2022). Replication of the natural selection of bad science. https://doi.org/10.31222/osf.io/sjyp3 (see pp. 28, 29).

Ladeiras-Lopes, R., Clarke, S., Vidal-Perez, R., Alexander, M., Lüscher, T. F., & On behalf of the ESC (European Society of Cariology) Media Committee and European Heart Journal. (2020). Twitter promotion predicts citation rates of cardiovascular articles: A preliminary analysis from the ESC journals randomized study. *European Heart Journal*, *41*(34), 3222–3225. https://doi.org/10.1093/eurheartj/ehaa211 (see p. 24).

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PLOS ONE*, *16*(6), 1–20. https://doi.org/10.1371/journal.pone.0253129 (see p. 23).

Luc, J. G. Y., Archer, M. A., Arora, R. C., Bender, E. M., Blitz, A., Cooke, D. T., Hlci, T. N., Kidane, B., Ouzounian, M., Varghese, T. K., & Antonoff, M. B. (2021). Does tweeting improve citations? one-year results from the TSSMN prospective randomized trial. *The Annals of Thoracic Surgery*, *111*(1), 296–300. https://doi.org/10.1016/j.athoracsur.2020.04.065 (see p. 24).

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology*, *16*(1), 274. https://doi.org/10.1186/s13059-015-0850-7 (see p. 20).

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, *5*, e16800. https://doi.org/10.7554/eLife.16800 (see pp. 20, 21).

Mikki, S. (2017). Scholarly publications beyond paywalls: Increased citation advantage for open publishing. *Scientometrics*, *113*(3), 1529–1538. https://doi.org/10.1007/s11192-017-2554-0 (see p. 20).

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021 (see p. 20).

National Academies of Sciences, Engineering, and Medicine. (2019). Improving reproducibility and replicability. In *Reproducibility and replicability in science* (pp. 105–142). National Academies Press (US). (See p. 20).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716 (see p. 20).

Pashler, H., & Wagenmakers, E.-J. (2012). Editor's introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253 (see p. 20).

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research worklow: Better science but more work. *Royal Society Open Science*, *9*, 211997. https://doi.org/10.1098/rsos.211997 (see p. 22).

Schneider, J., Rosman, T., Kelava, A., & Merk, S. (2020). (Re)Building trust? journals' open science badges influence trust in scientists. https://doi.org/10.23668/PSYCHARCHIVES.3364 (see p. 20).

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632 (see p. 20).

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384. https://doi.org/10.1098/rsos.160384 (see pp. 20, 21, 22, 24, 29).

Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science. In *Stevens' handbook of experimen-*

tal psychology and cognitive neuroscience (pp. 1–47). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119170174.epcn519 (see p. 20).

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*, 553–571. https://doi.org/10.1007/s10869-020-09695-3 (see pp. 22, 23).

Vazire, S. (2020). Open scholarship: Where are the self-correcting mechanisms of science? Retrieved December 12, 2022, from https://www.google.com/search?q=simine+vazire+opening+the+hood&oq=simine+vazire+opening+the+hood+&aqs=chrome..69i57j33i160l2.4891j0j7&sourceid=chrome&ie=UTF-8#fpstate=ive&scso=_ouyWY8G6NKWR9u8P5fSGqAo_31:0&vld=cid:4a209471,vid:Vfc98WDfDJE,st:752 (see p. 29).

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078 (see p. 20).

| Appendices

A. Sensitivity Analyses for Lab Proportions and Characteristics

For all sensitivity analyses, we did not differentiate between simulation runs with two qualitatively different patterns of characteristics (see Results section). Instead, we averaged over all simulation runs. Each of the following Figures contains various parameter combinations. One additional parameter (i.e., the payoff advantage for advocating OS γ) differentiates between Figures. Figures 5, 6, and 7 show the lab proportions for the first set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively; Figures 8, 9, and 10 show the community characteristics for the first set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively. Similarly, Figures 11, 12, and 13 show the lab proportions for the second set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively; Figures 14, 15, and 16 show the community characteristics for the second set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively.

Figures 5, 6, 7, 11, 12, and 13 clearly demonstrate that the lab proportions are robust against specific choices of parameter combinations. In all cases, the "practice; advocate" lab type wins and suppresses the other lab types. Although there is more variation in the community characteristics for different parameter combinations (see Figures 8, 9, 10, 14, 15, and 16), the overall trends are still quite robust.
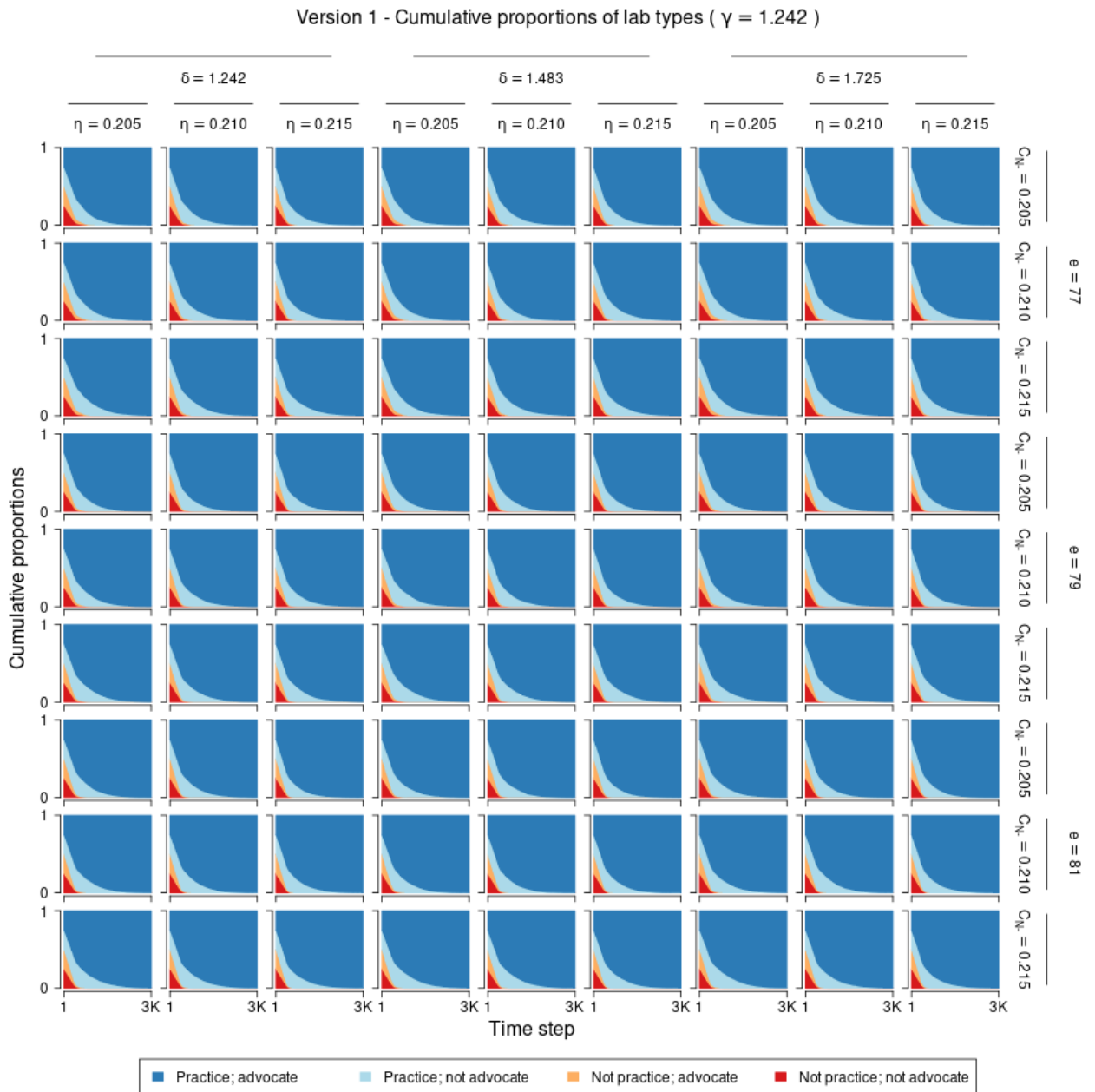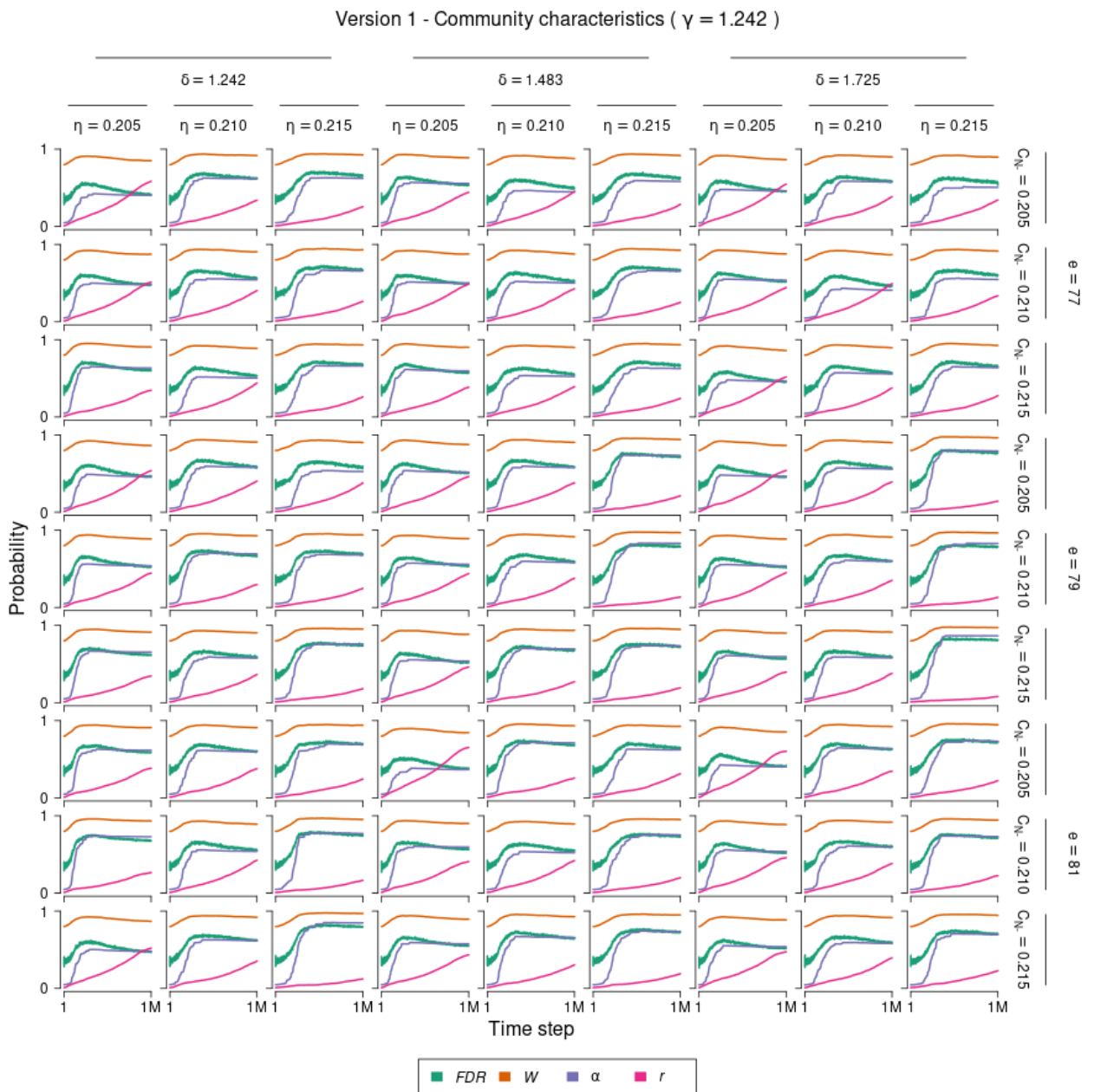
Figure 5 Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the first set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
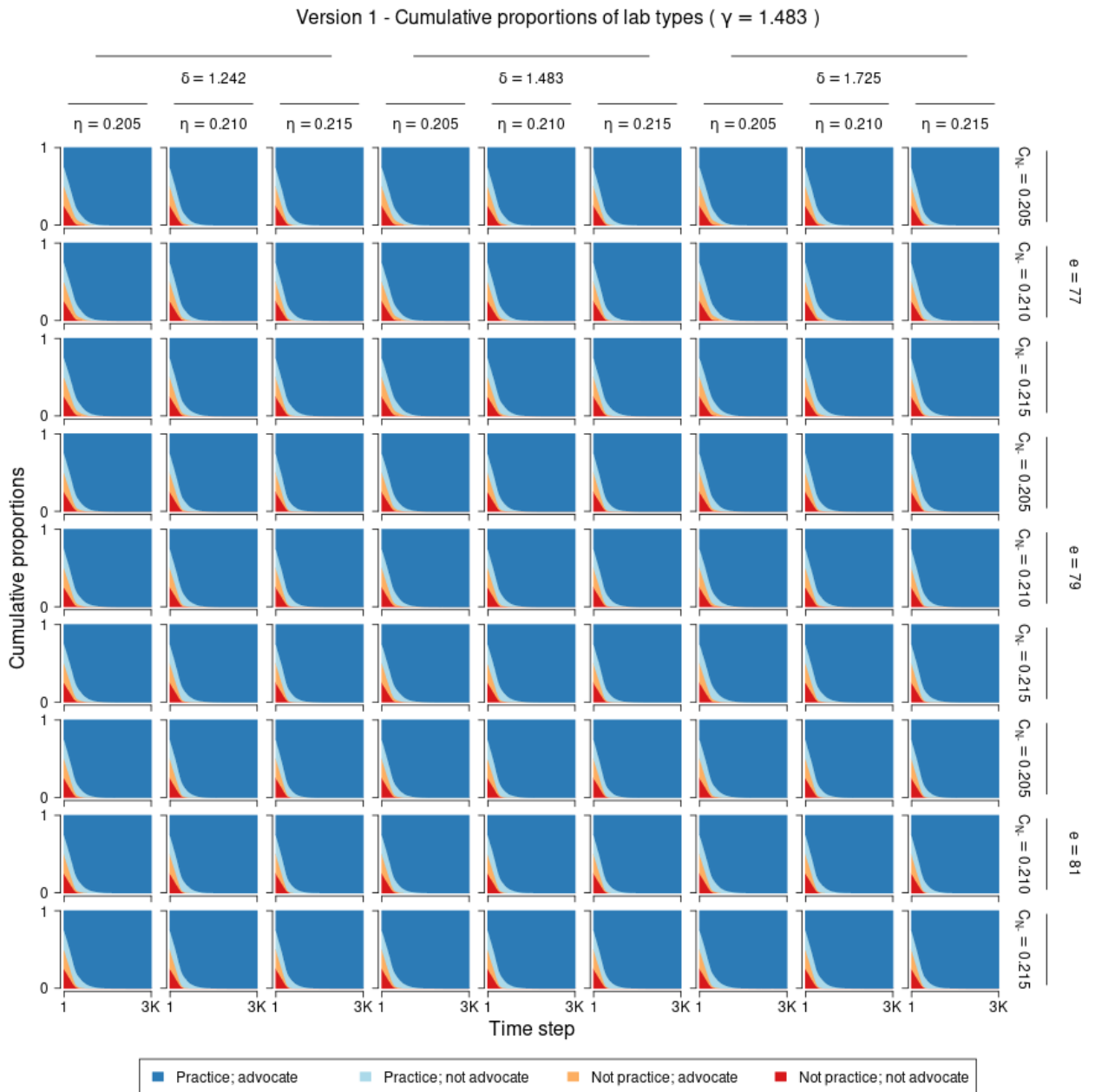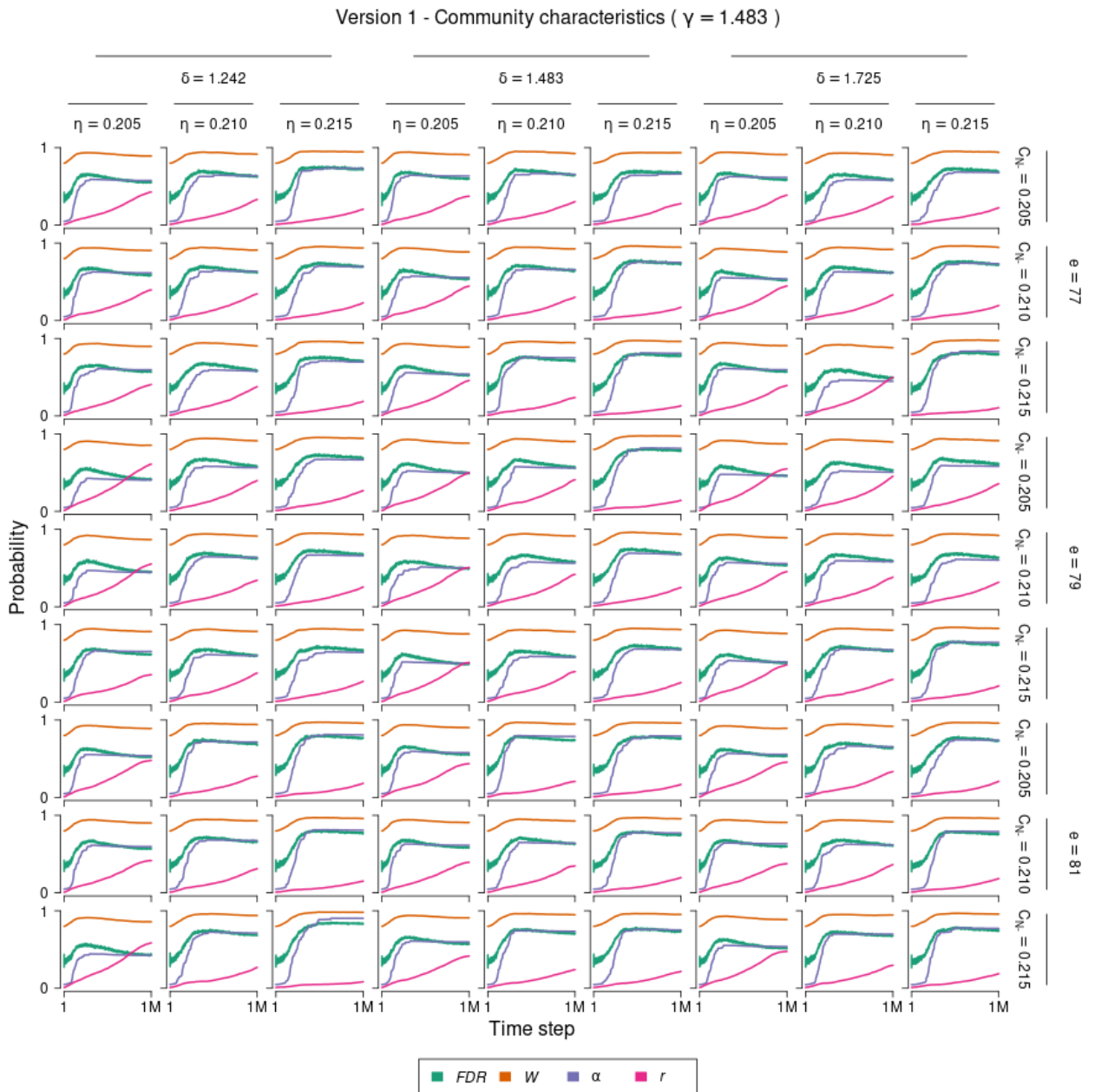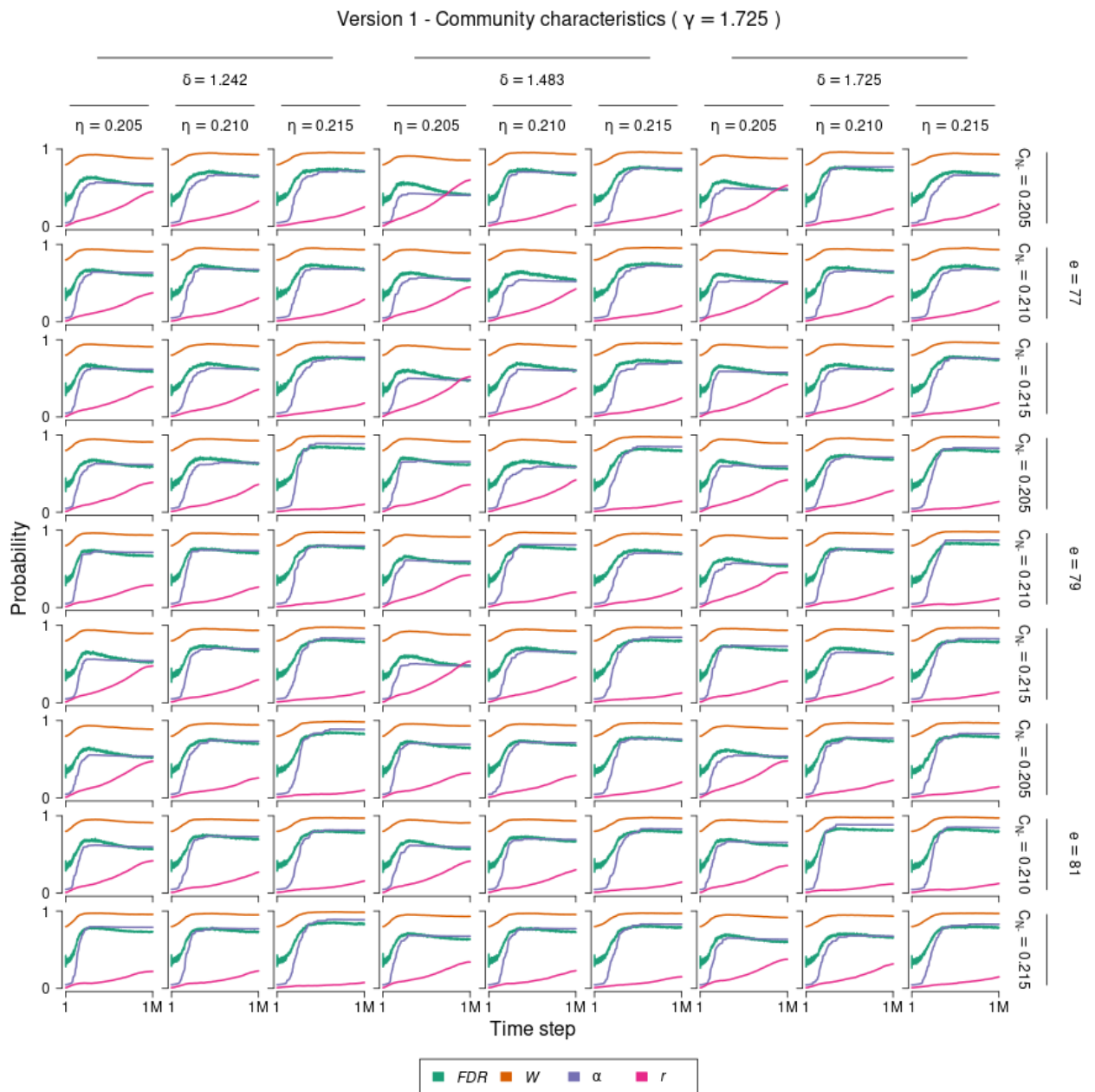
Figure 6 Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the first set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
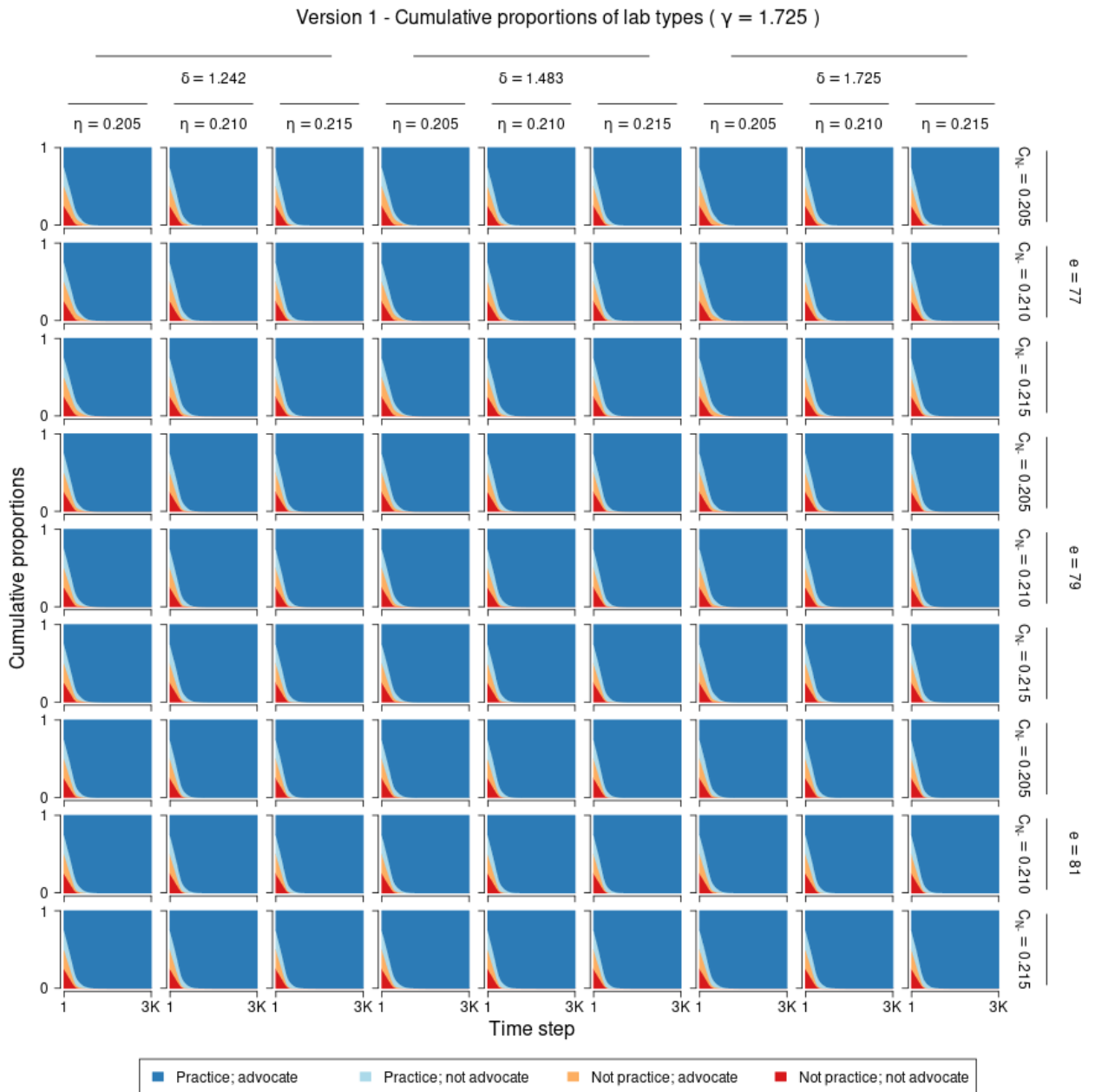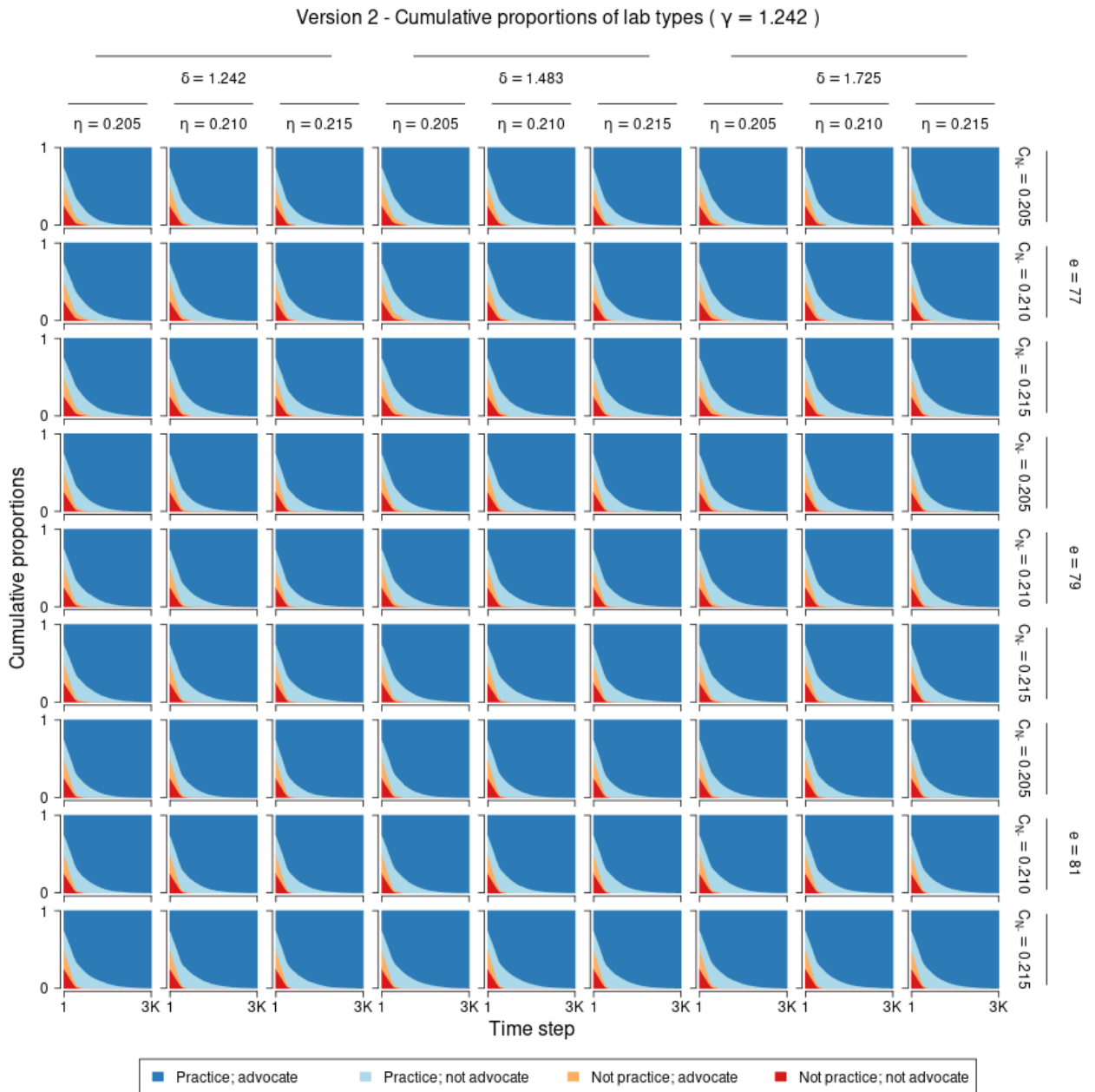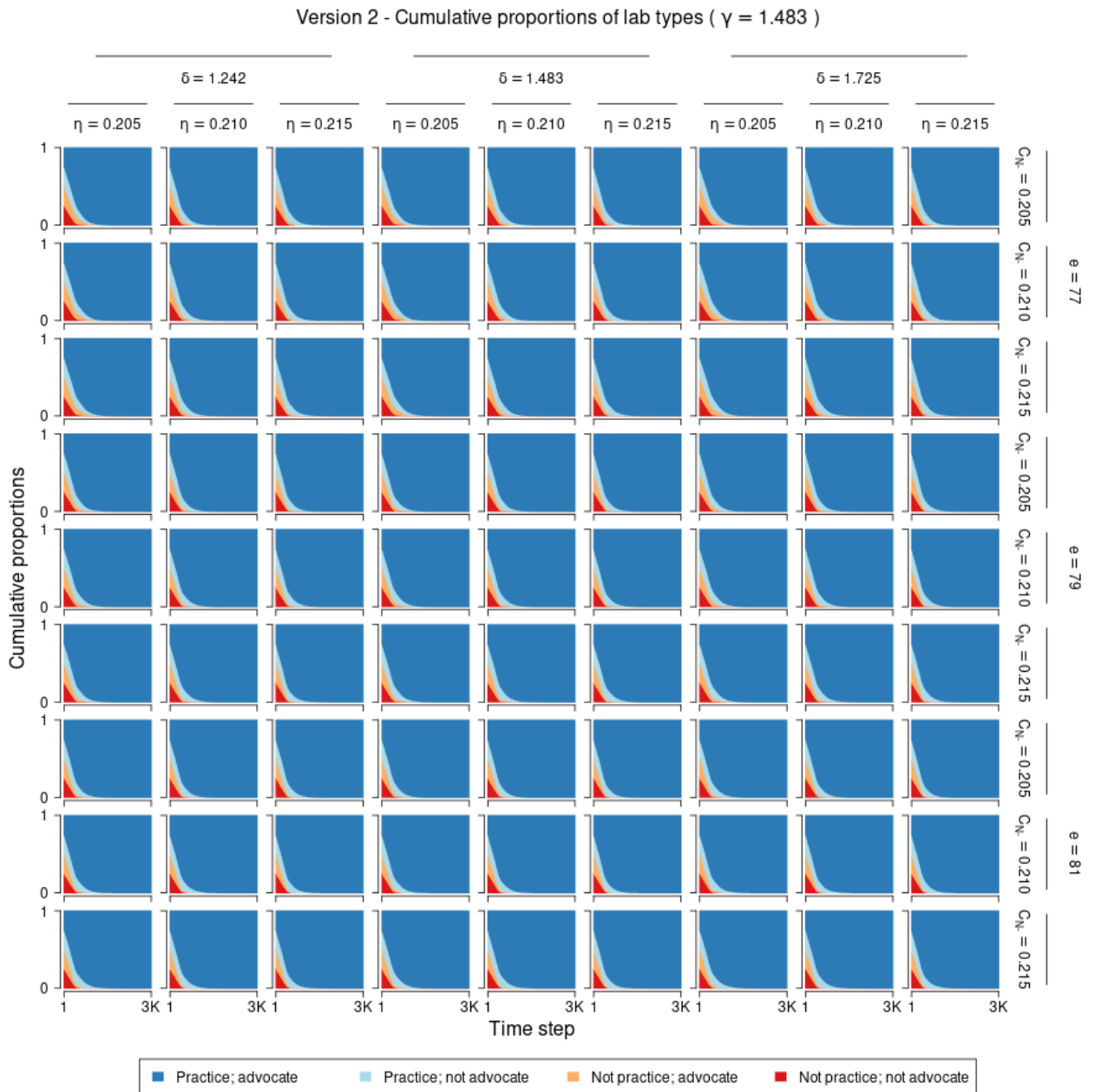
Figure 7 Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the first set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

Figure 8   Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the first set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
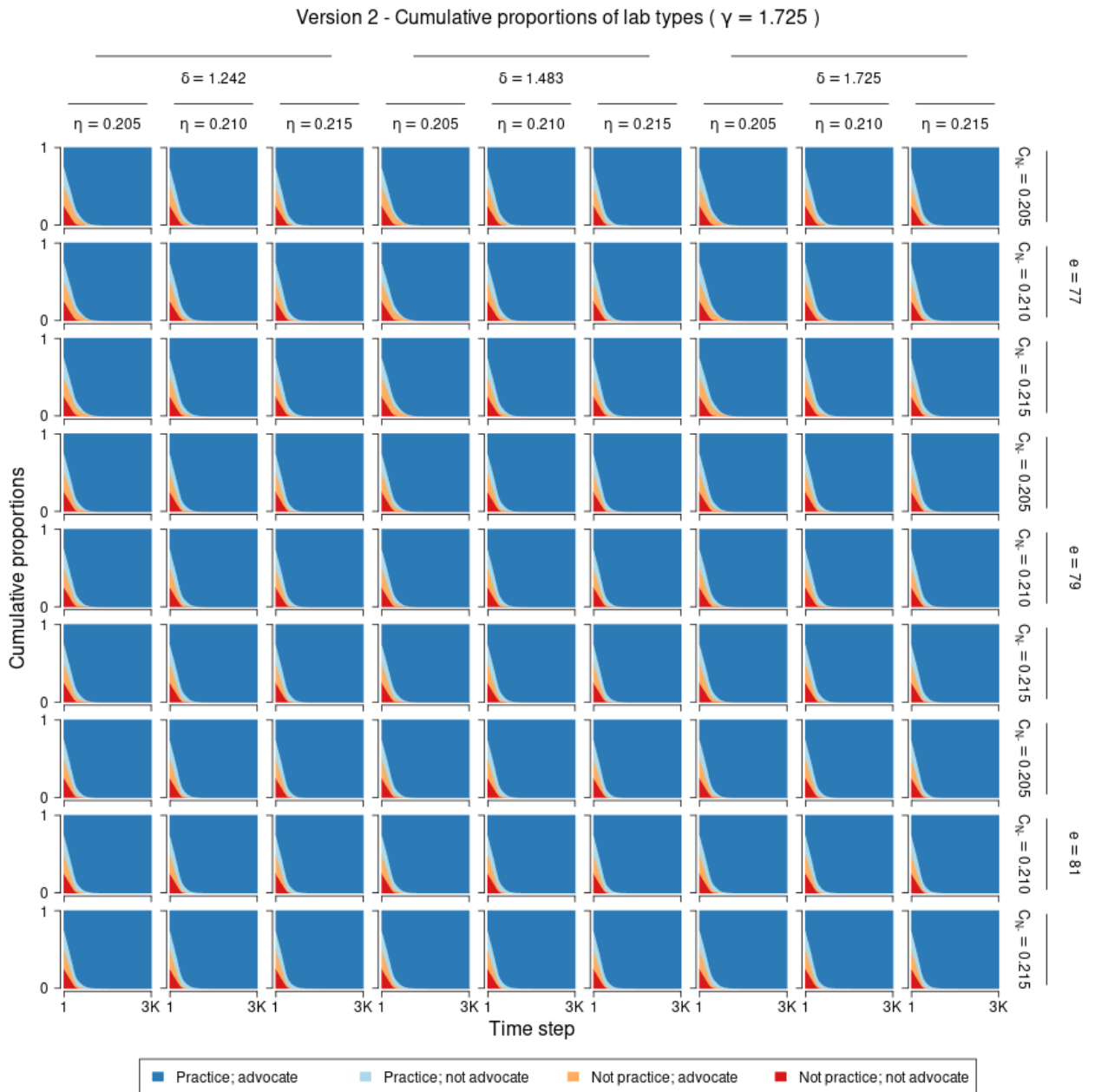
Figure 9 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the first set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

Figure 10 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the first set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
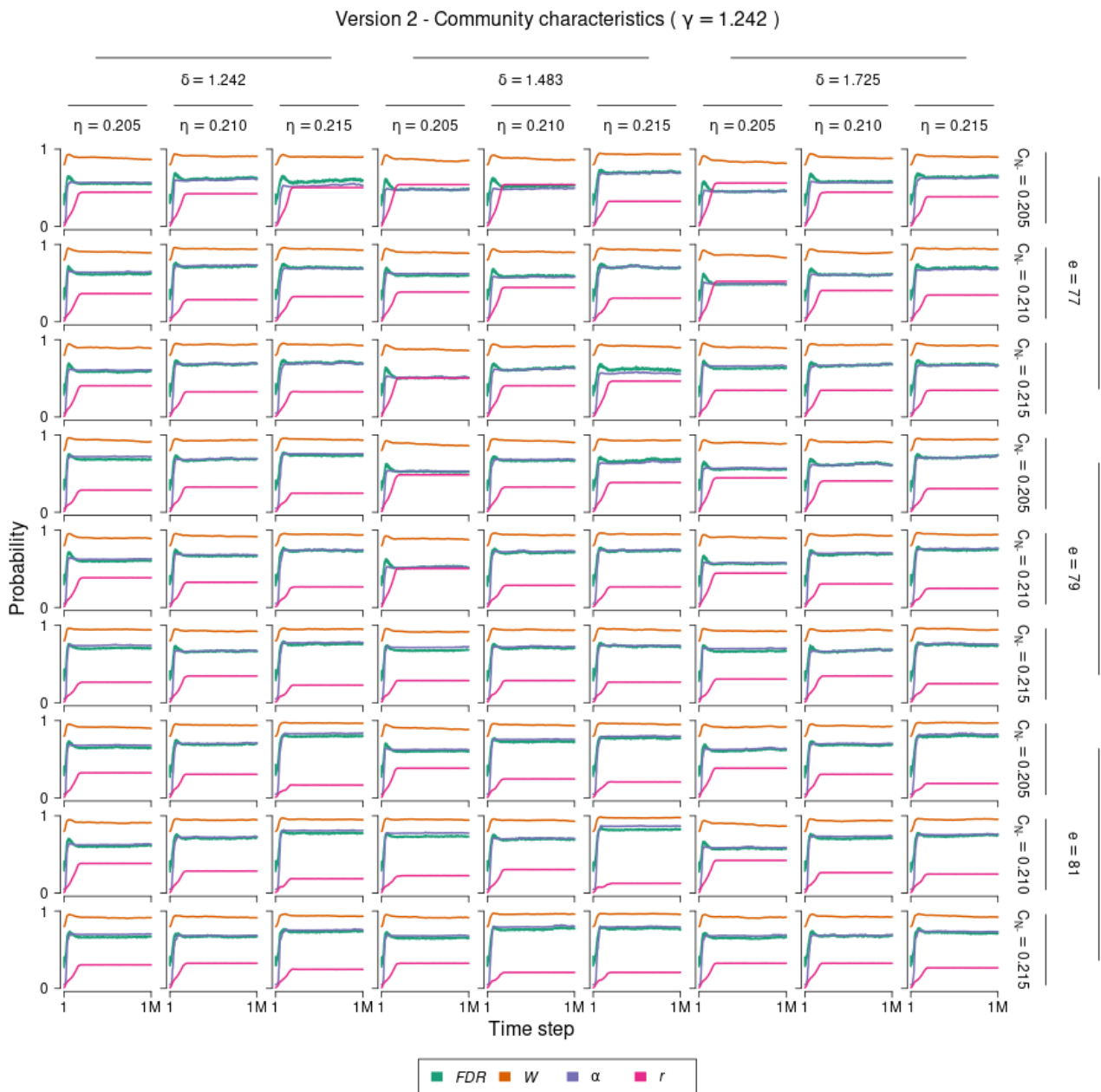
Figure 11 Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the second set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
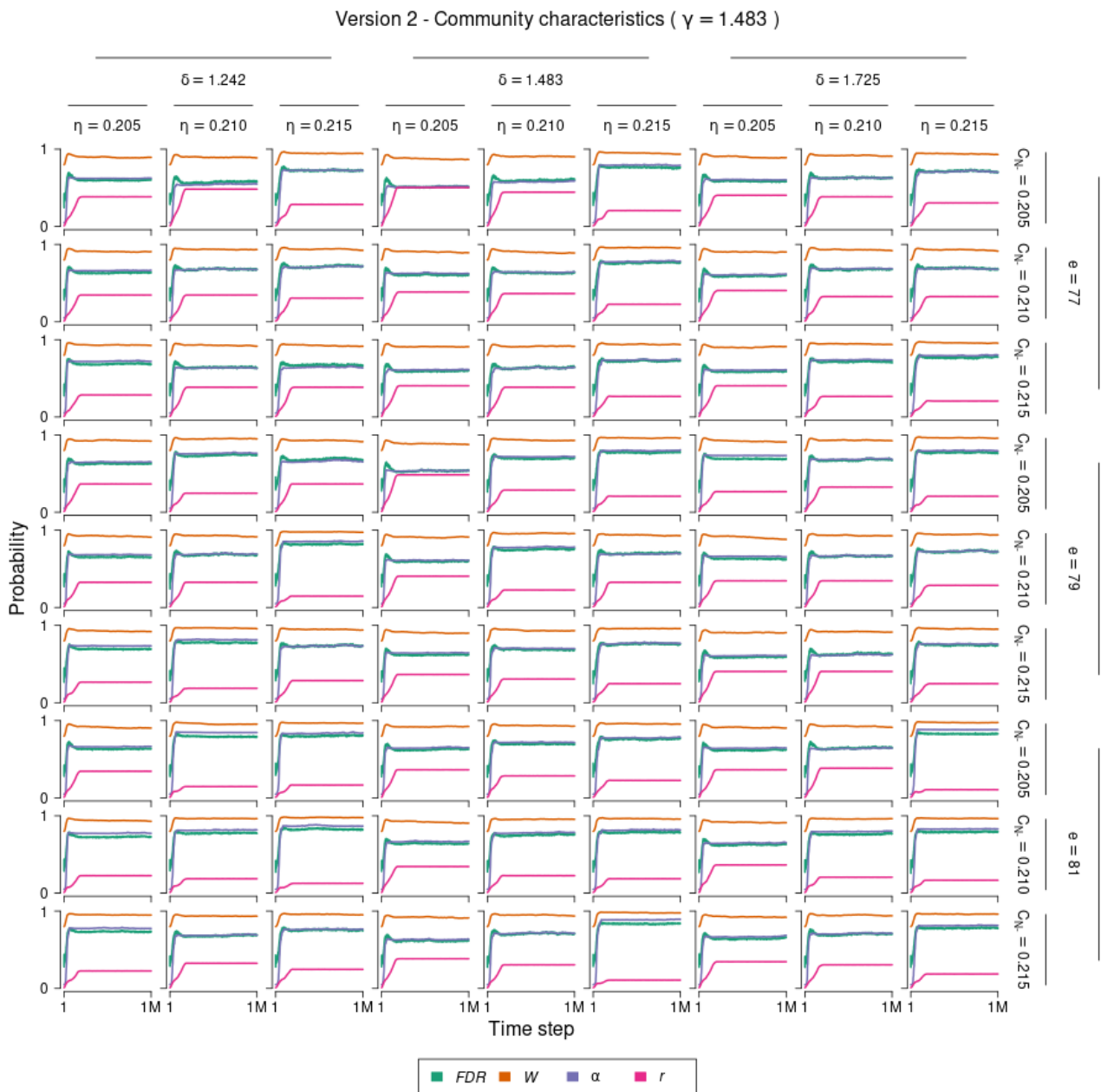
Figure 12 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the second set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

Figure 13 Cumulative proportions of lab types during the first $3,000$ out of $1,000,000$ time steps of the second set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
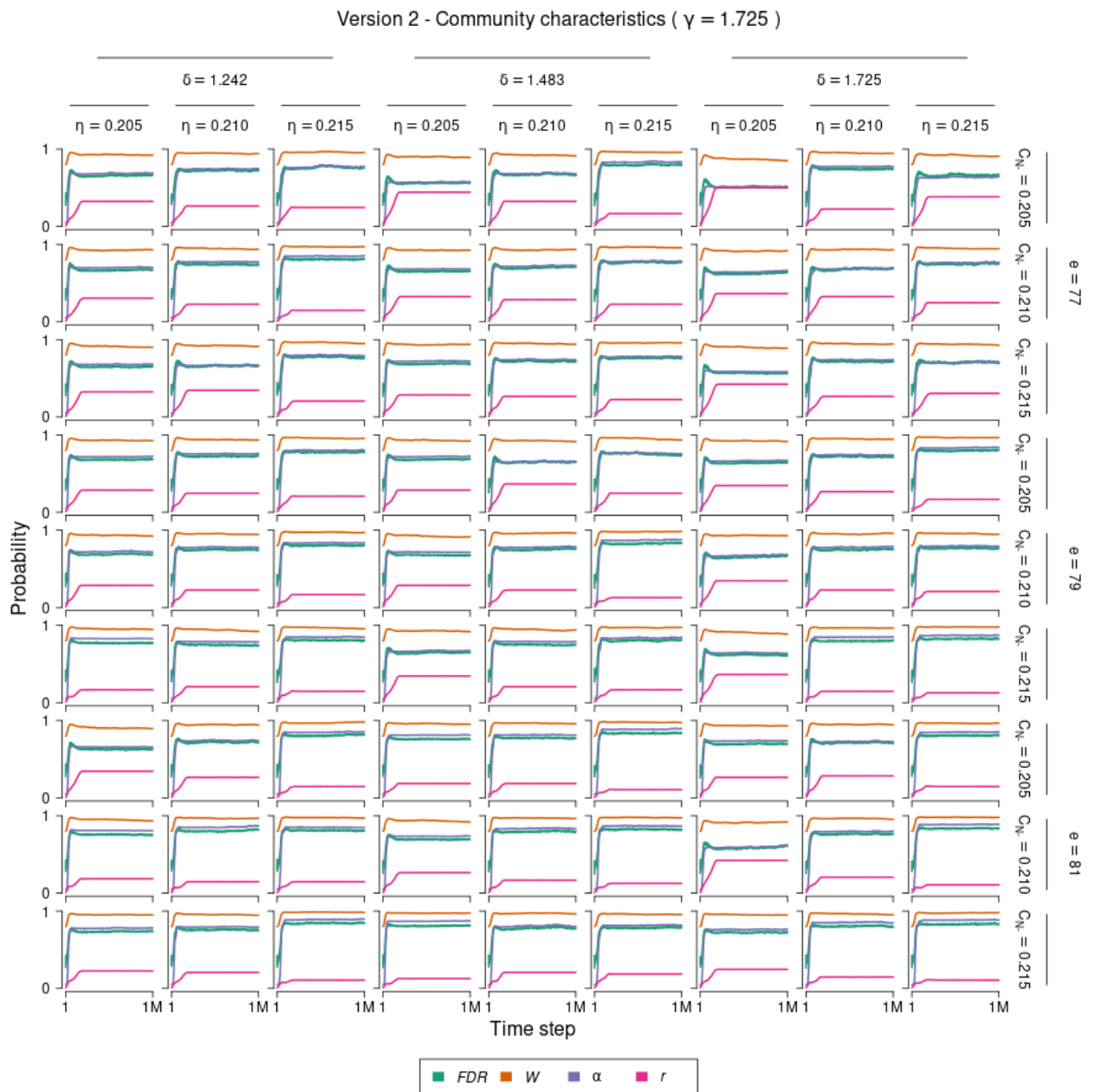
Figure 14 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the second set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.
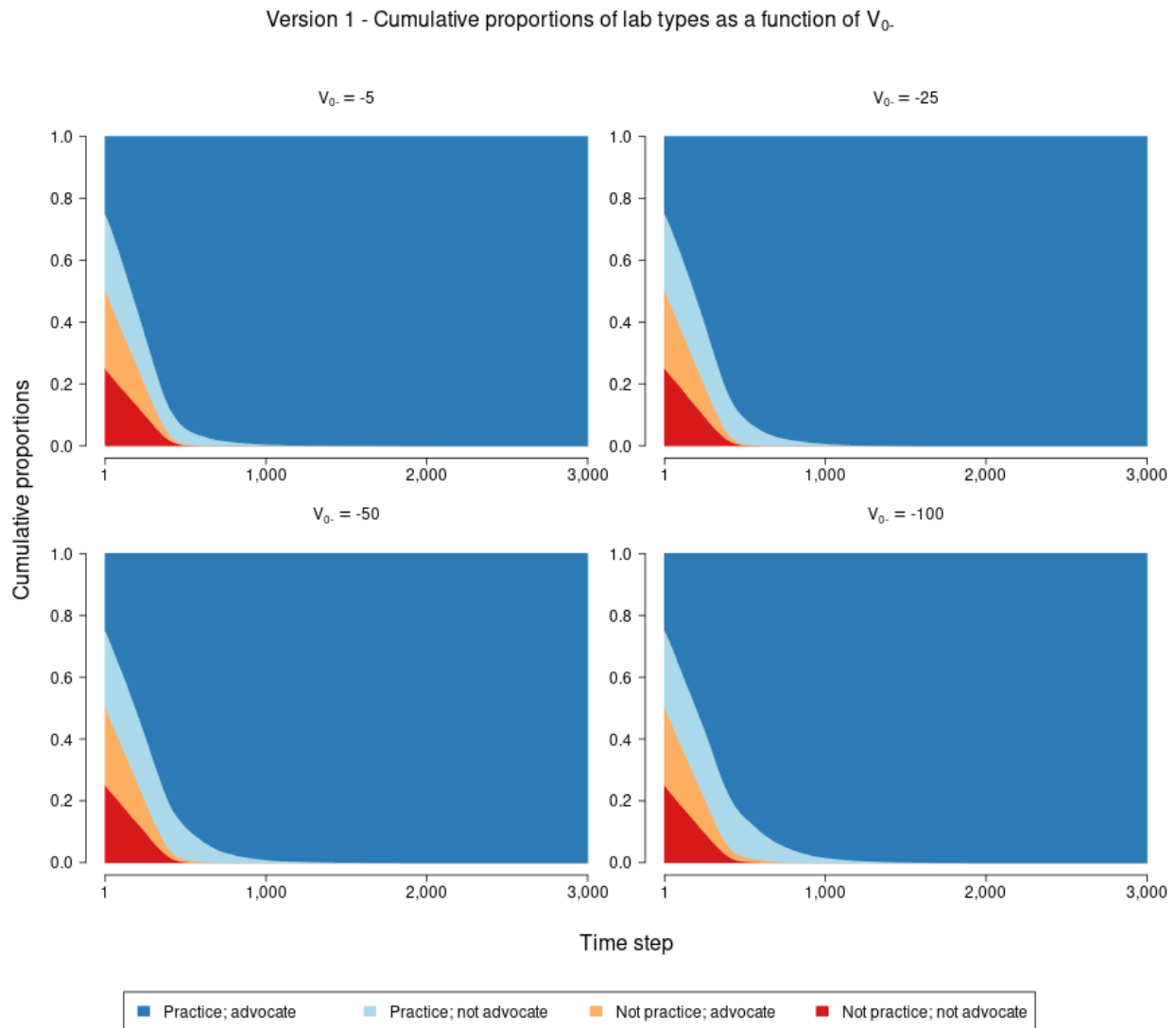
Figure 15 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the second set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

Figure 16 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the second set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

B. Lab proportions and characteristics as a function of V_{0-}

For these additional sensitivity analyses, we did not differentiate between simulation runs with two qualitatively different patterns of characteristics (see Results section). Instead, we averaged over all simulation runs. Figure 17 shows the lab proportions for the first set of simulations as a function of V_{0-} and Figure 18 shows the community characteristics for the first set of simulations as a function of V_{0-}. All parameter values, except for V_{0-} are fixed at the values that were used in the main analyses (see Table 2). Figure 17 clearly demonstrates that the lab proportions are robust against specific choice of V_{0-}. Figure 18 shows that the development of FDR, W, and α is also robust against variations of V_{0-}; all of them increase quickly and remain at a high level. However, it can be seen that r increases more strongly the higher the value for V_{0-}.

Figure 17   Cumulative proportions of lab types as a function of V_{0-} during the first $3,000$ out of $1,000,000$ time steps of the first set of simulations. The panels correspond to different values of V_{0-}. All other parameter values are fixed at the values that were used in the main analyses (see Table 2). Note that panels show an average of all 50 simulation runs.
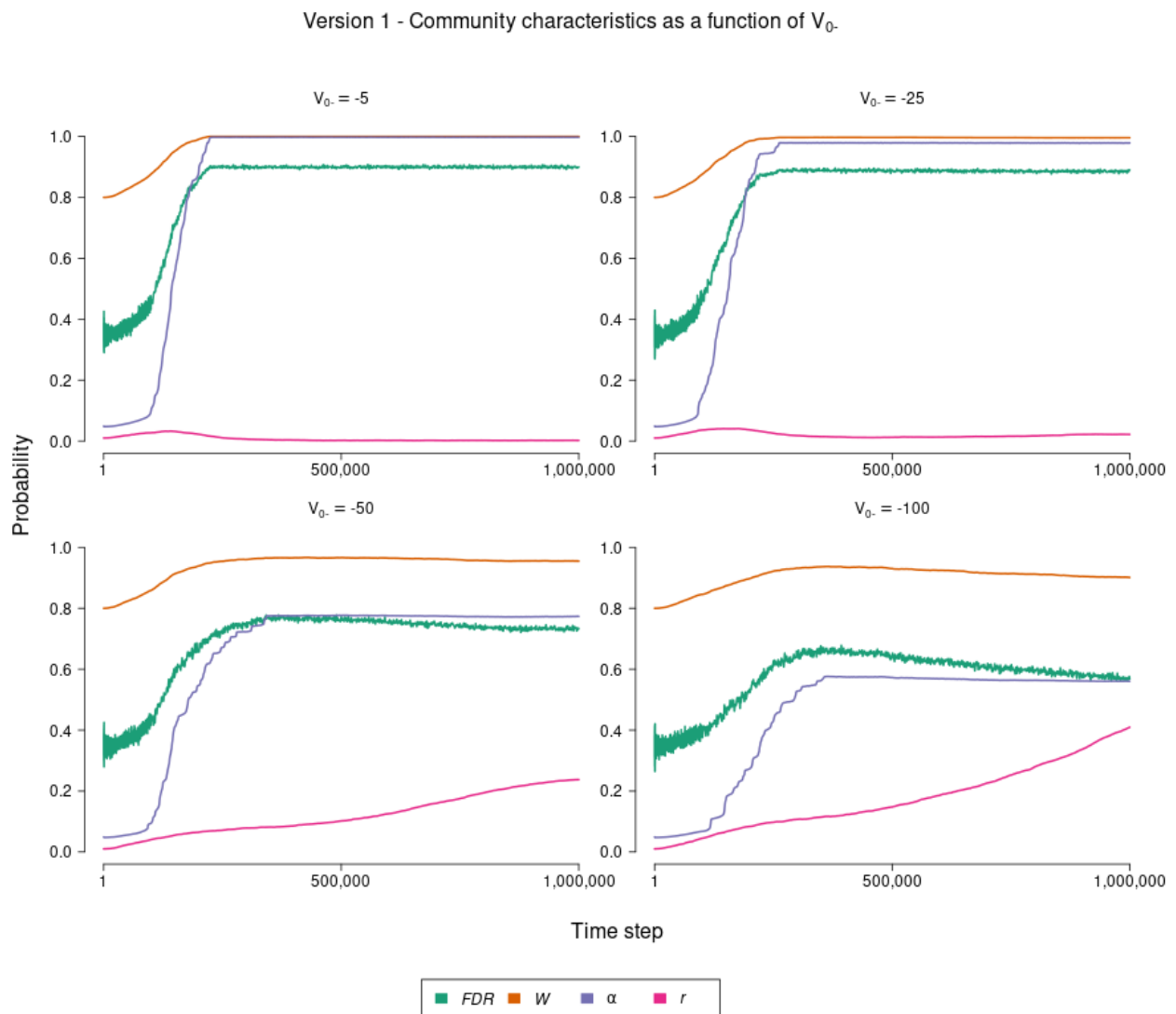
Figure 18 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs as a function of V_{0-} over all $1,000,000$ time steps of the first set of simulations. The panels correspond to different values of V_{0-}. All other parameter values are fixed at the values that were used in the main analyses (see Table 2). Note that panels show an average of all 50 simulation runs.