PLOS ONE

RESEARCH ARTICLE

Replication, Communication, and the Population Dynamics of Scientific Discovery

Richard McElreath[1,2]*, **Paul E. Smaldino**[1]

1 Department of Anthropology, University of California Davis, Davis, CA, United States of America, **2** Center for Population Biology, University of California Davis, Davis, CA, United States of America

* mcelreath@ucdavis.edu

Abstract

Many published research results are false (Ioannidis, 2005), and controversy continues over the roles of replication and publication policy in improving the reliability of research. Addressing these problems is frustrated by the lack of a formal framework that jointly represents hypothesis formation, replication, publication bias, and variation in research quality. We develop a mathematical model of scientific discovery that combines all of these elements. This model provides both a dynamic model of research as well as a formal framework for reasoning about the normative structure of science. We show that replication may serve as a ratchet that gradually separates true hypotheses from false, but the same factors that make initial findings unreliable also make replications unreliable. The most important factors in improving the reliability of research are the rate of false positives and the base rate of true hypotheses, and we offer suggestions for addressing each. Our results also bring clarity to verbal debates about the communication of research. Surprisingly, publication bias is not always an obstacle, but instead may have positive impacts—suppression of negative novel findings is often beneficial. We also find that communication of negative replications may aid true discovery even when attempts to replicate have diminished power. The model speaks constructively to ongoing debates about the design and conduct of science, focusing analysis and discussion on precise, internally consistent models, as well as highlighting the importance of population dynamics.

Citation: McElreath R, Smaldino PE (2015) Replication, Communication, and the Population Dynamics of Scientific Discovery. PLoS ONE 10(8): e0136088. doi:10.1371/journal.pone.0136088

Editor: Daniele Marinazzo, Universiteit Gent, BELGIUM

Received: May 1, 2015

Accepted: July 29, 2015

Published: August 26, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The Division of Social Sciences Dean's Office at the University of California Davis provided financial support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Imagine two of your close colleagues have just heard about attempts to replicate their positive research findings. Colleague A is thrilled that the attempt was successful. Colleague B is upset that the attempt was unsuccessful. What is the probability that Colleague A's hypothesis is true? What is the probability that Colleague B's hypothesis is false?

This is not a fair quiz, because in truth no one knows the answers to these questions. The absence of replication in many fields [2–4], combined with the absence of a formal framework for understanding replication, makes it difficult to even outline an answer. In the absence of replication, there is substantial concern that many published findings may be false [1], an

argument with empirical support [5–7]. The history of science buttresses these observations. A recent catalog of false discoveries of chemical elements outnumbers the current number of real elements in the periodic table [8]. In addition to concerns about replication are concerns about research practice and publication bias. Without knowing how many studies were conducted but not published, it is not possible to assign evidential value to either initial findings or replications. And it is not yet easy to acquire empirical evidence about these factors, as even the best empirical studies of publication bias still rely upon researcher self-report [3].

Thus many opinions can be sustained about the evidential value of both initial findings and replications. As a result, recent controversies over failed replications demonstrate a lack of consensus on norms for replication and publication [9–12]. What is the evidential value of replication, positive or negative? What is the impact of publication bias [13]? If replication is part of an "invisible hand"[14] that corrects scientific errors, how much replication is needed? And what are the risks of poorly designed or interpreted replication attempts [9]? When replication is not possible or practical, what other measures can be taken to improve the reliability of research?

These questions remind us that little is understood about the population dynamics of discovery, replication, and scientific communication. Much more attention has been given to individual methods of research design and data analysis. And while it is useful to analyze research methods in isolation, such calculations are unsatisfying. A lot of research activity is hidden from the public record. This means the actual number of findings for an hypothesis may never be known [13]. And since researchers select hypotheses for further study from the literature itself, findings and publication biases cascade into other findings, interacting with biases and incentives [15].

To know the evidential value of research, we must study the population dynamics that produce it [14, 16–18]. So here we construct and solve a mathematical model of scientific beliefs formed by a population of boundedly rational agents who accumulate evidence for and against hypotheses. We adopt a general signal detection framework that may apply to diverse statistical paradigms, whether *p*-valued or Bayesian. We study the joint dynamics that arise from replication, publication bias, and differences in research quality between original studies and replications. Our goal is not to accurately simulate science, but rather to understand it better using the same reductionist tools that have been so successful in illuminating population dynamics more generally [19, 20]. Our model implicitly provides, for example, a neutral model of scientific dynamics in which all hypotheses are false and yet discoveries are continuously published. It also provides a range of "selectionist" models that might be compared to data. The clarity of a quantitative framework will stimulate and clarify the development of later empirical investigation and experimental intervention.

The paper proceeds by first outlining the dynamic structure of the model. We then solve the model for both its long-run dynamics and its epistemological implications—what should a rational agent believe about an hypothesis, given a record of published results? We present a general interpretation of the joint dynamics, so the reader can extrapolate lessons from our simple model to the complexity and diversity of real science. We conclude by relating our results to ongoing debates about improving the reliability of scientific research.

Model Description

The model is illustrated in Fig 1. We have also constructed an interactive, web-based tutorial on the conceptual foundations of the model, as well as fully adjustable simulation code, available at http://xcelab.net/replication/. A population of researchers studies many different hypotheses. Each hypothesis is either *true* (green) or *false* (red). These hypotheses could be

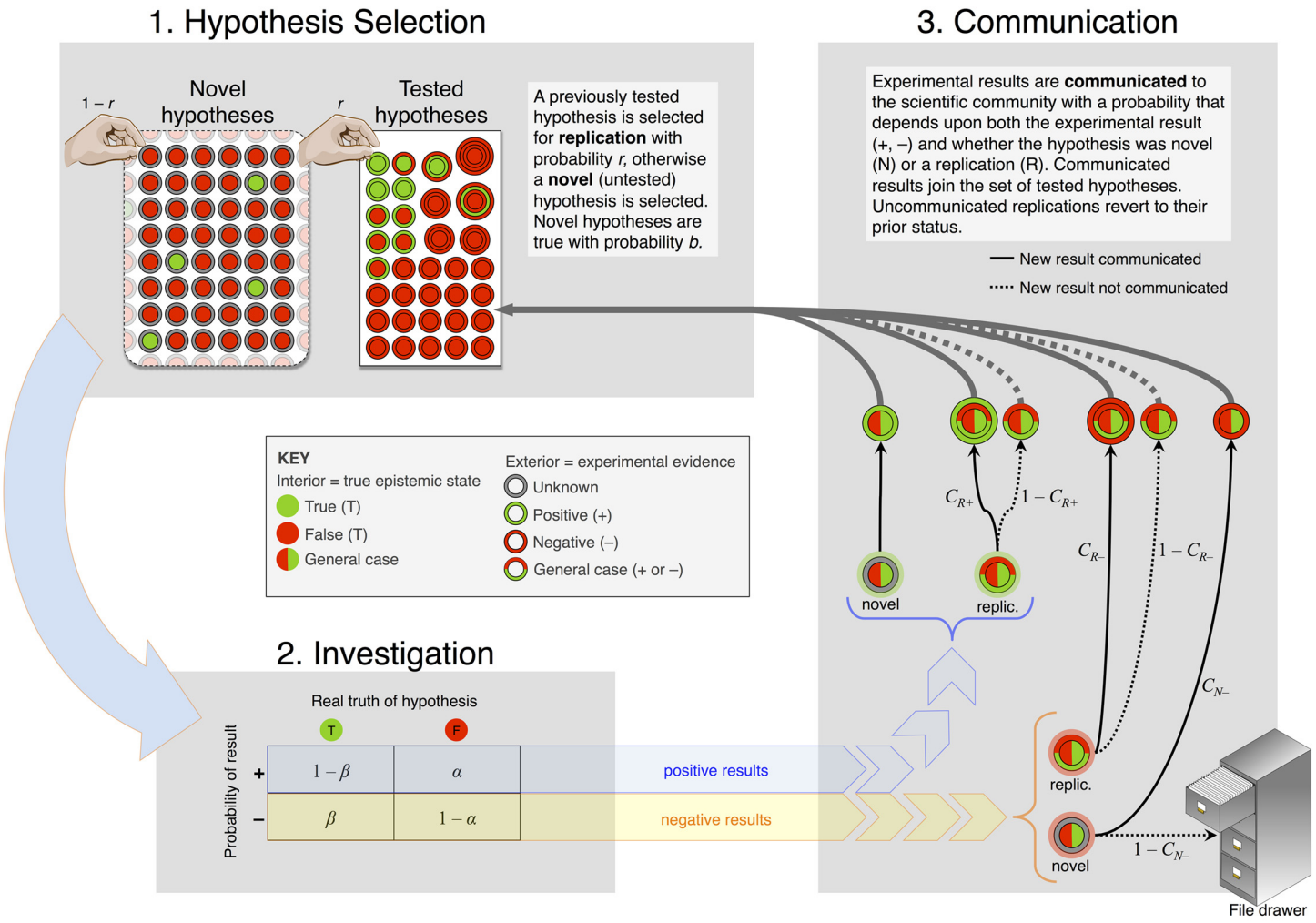

Fig 1. Population dynamics of replication.

doi:10.1371/journal.pone.0136088.g001

simple associations, such as *green jelly beans cause acne* [21], or more general claims, such as *evolution is predictable*. Research results in either a *positive* or a *negative* finding. These findings may be the result of formal hypothesis tests or informal assessments. True hypotheses produce positive findings more often than do false hypotheses, but the researchers never know for sure which hypotheses are true. Under these assumptions, the only information relevant for judging the truth of an hypothesis is its *tally*, the difference between the number of published positive findings and the number of published negative findings for each hypothesis, and we summarize results in terms of these tallies. In reality, much other information is relevant to judging the truth of an hypothesis. Our assumptions are tactical ones. More complex models of scientific communication are possible, but any such model must include the components in our model, and so our results establish a critical baseline.

Each time interval, research activity has three stages that alter these tallies. In stage 1 (Fig 1, upper-left) each researcher chooses to investigate one of n previously published hypotheses,

with probability r, or a novel hypothesis, with probability $1 - r$. When replicating, a researcher chooses a previously published hypothesis at random and performs a new study of it. Later, we allow researchers to target hypotheses with specific tally values, rather than choosing at random. A novel hypothesis is true with probability b, the *base rate*, reflecting mechanisms of hypothesis formation. Untutored intuition, for example, may be expected to yield a very low b. Genome wide association studies likewise have low b, because relatively few loci are associated with any particular phenotype. There is no consensus on base rate, except that most scientists we know believe their own personal b values are better than average. So we allow b to vary freely in the model.

In stage 2, a true hypothesis produces a positive finding $1 - \beta$ of the time, its *power*. A false hypothesis produces a positive finding α of the time, its *false positive rate*. We assume that $1 - \beta > \alpha$. Later we allow the values of β and α to differ between replication attempts and initial studies. Note that β and α are not merely properties of a statistical procedure, but rather of an entire investigation. For example, using several procedures and selecting the one that produces a positive result will inflate α [22].

In stage 3, findings may be communicated to other researchers. Not every finding is communicated, either because no one tries to communicate it or rather because it cannot be published. Only communicated findings can adjust a tally. Let $c_{\mathrm{N}\text{-}}$ be the probability that a negative (−) finding about a new (N) hypothesis is communicated. We assume for simplicity that all new positive results are communicated ($c_{\mathrm{N}+} = 1$). Even though replication findings are evidentially equivalent to novel findings, they may be communicated with different probability. Let $c_{\mathrm{R}\text{-}}$ and $c_{\mathrm{R}+}$ be the probabilities that replications with negative and positive findings, respectively, are communicated.

These assumptions define the dynamics of the expected numbers of true and false hypotheses with a given tally. We present the full recursions in S1 Text. In the simplest case (full communication: $c_{\mathrm{N}\text{-}} = c_{\mathrm{R}\text{-}} = c_{\mathrm{R}+} = 1$), the number $n_{\mathrm{T},s}$ of true hypotheses with an observed tally s in the next time step is given by:

$$n'_{\mathrm{T},s} = n_{\mathrm{T},s} + anr\left(-\frac{n_{\mathrm{T},s}}{n} + \frac{n_{\mathrm{T},s-1}}{n}(1-\beta) + \frac{n_{\mathrm{T},s+1}}{n}\beta\right) \quad (1)$$

where $a > 0$ is the rate of research activity as a proportion of n. This expression says that the number in the next time step is just the current number plus all of the flows in and out caused by replications. In the case that $s = -1$ or $s = 1$, there is an additional term $an(1 - r)b\beta$ or $an(1 - r)b(1 - \beta)$, respectively, to represent the inflow of novel findings. Recursions $n'_{\mathrm{F},s}$ for false hypotheses are constructed from a change in variables: $1 - \beta \rightarrow \alpha$, $b \rightarrow 1 - b$. Notice that this implies that the model is easily extended to any number of hypothesis types, such as effect size differences, that differ in power and false-positive rate. We analyze the *true*/*false* dichotomy because of its prominence and simplicity.

Analysis

By literature review, a tally can be constructed for any given hypothesis. Given an observed tally, but a number of possibly unobserved studies, what is the probability that an hypothesis is correct? The model allows us to address this question for a diversity of scenarios. Before presenting the solutions, note that the answers that the model provides can be understood both from a pure population dynamics perspective and from a probabilistic reasoning perspective. From the dynamics perspective, the population will converge from any initial condition to a unique steady state in which the solutions give *frequencies* of true hypotheses at each tally value. Equally valid is the epistemological perspective that the solutions tell us for any unique

hypothesis the *probability* it is true, given a state of information [23]. One consequence of this is that the solutions do not require that all hypotheses share the same parameter values.

For each tally value s, we solved for the steady state proportions of true and false hypotheses, $\hat{p}_{\mathrm{T},s}$ and $\hat{p}_{\mathrm{F},s}$. We also derived the same solutions under the probabilistic interpretation, and verified our solutions numerically and through stochastic simulation. We present complete analytical solutions in S1 Text. In the simplest case (for full communication), solutions take the form:

$$\hat{p}_{\mathrm{T},s} = b(1-r)\sum_{m=1}^{\infty} r^{m-1}\binom{m}{\frac{1}{2}(m+s)}(1-\beta)^{\frac{1}{2}(m+s)}\beta^{\frac{1}{2}(m-s)} \tag{2}$$

This expression defines an infinite geometric series of binomial probabilities arising from all of the different possible histories by which a true hypothesis could achieve a tally of s, for every possible number of findings m. In the majority of cases, only the first few terms of the series are important, because of the leading factor r^{m-1}. This fact also informs us that the rate of convergence to steady state will be quite rapid, unless r is large.

For any particular tally, for example $s = 1$, expression (2) yields a closed-form solution like:

$$\hat{p}_{\mathrm{T},1} = \frac{b(1-r)}{2\beta r^2}\left((1-4r^2\beta(1-\beta))^{-\frac{1}{2}} - 1\right) \tag{3}$$

For arbitrary communication parameters, the solutions have a similar structure, but are instead a series of multinomial probabilities in which the events are combinations of findings (+ or −) and communication outcomes.

These solutions are not easy to interpret by inspection. But they do provide answers to the question: *what is the probability that an hypothesis with a given tally is correct?* For any tally s, we can calculate:

$$\Pr(\mathrm{true}|s) = \frac{\hat{p}_{\mathrm{T},s}}{\hat{p}_{\mathrm{T},s} + \hat{p}_{\mathrm{F},s}}, \qquad \Pr(s|\mathrm{true}) = \frac{\hat{p}_{\mathrm{T},s}}{\sum_i \hat{p}_{\mathrm{T},i}}, \qquad \Pr(s|\mathrm{false}) = \frac{\hat{p}_{\mathrm{F},s}}{\sum_i \hat{p}_{\mathrm{F},i}} \tag{4}$$

The *precision* of a tally s is Pr(true|s), the proportion of hypotheses with tally s that are true. The *sensitivity*, Pr(s|true), is the proportion of true hypotheses with tally s. It indicates where the true hypotheses are. Sensitivity is important because a high precision for a tally s is little help when there are few hypotheses that achieve a tally s. And the *specificity*, Pr(s|false), is the proportion of false hypotheses with tally s, indicating where the false hypotheses are. We use these definitions to explain the behavior of the system.

Overall dynamics

Fig 2 describes the overall dynamics of precision, as a function of the different parameters. In each panel, the trend lines show the proportion of true hypotheses at each tally on the vertical axis. The tally corresponding to each trend is indicated by a number. The horizontal axis in each panel varies a single parameter. Each vertical hairline shows the value of each parameter that is held constant in other panels. This figure is complex. We'll use it to highlight the most important factors in the reliability of findings and demonstrate counter-intuitive aspects of communication. Then in the next section, we'll turn to a more general explanation of the causes of these results.

There are two clusters of plots. The top cluster represents a normatively optimistic scenario, with an auspicious base rate ($b = 0.1$), unusually high power ($1 - \beta = 0.8$), low false-positive rate ($\alpha = 0.05$), and high communication rates. The bottom cluster represents a pessimistic, or perhaps more realistic [24, 25], scenario with low base rate ($b = 1/1000$), lower power ($1 - \beta =$

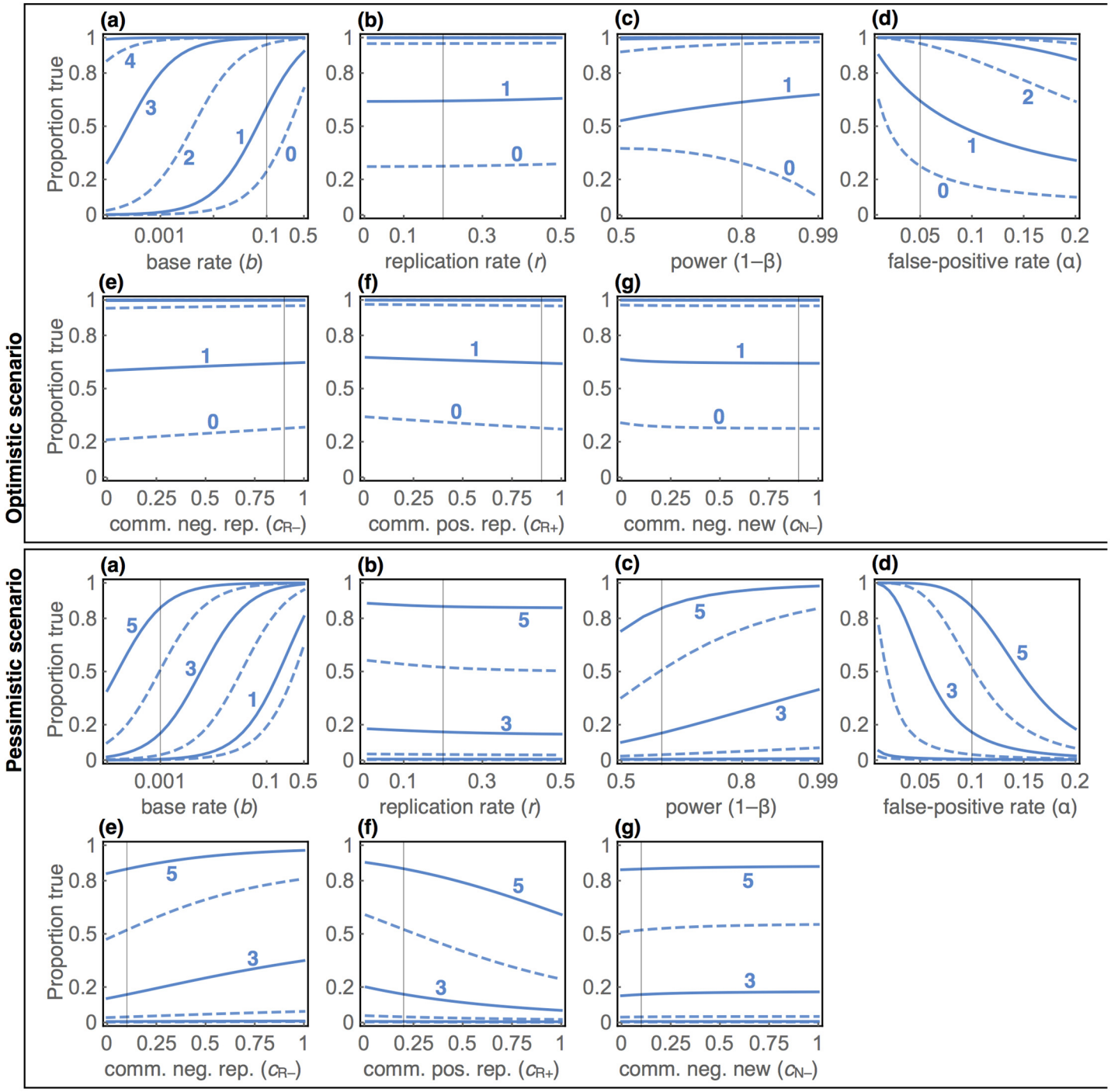

Fig 2. Effects of base rate, replication, power, false-positives, and communication on the probability that an hypothesis with a given tally is true. The two clusters illustrate difference scenarios. The blue trends, each labeled with its tally value, show precision as it varies by the parameter on each horizontal axis. The numbers indicate the tally of a curve. Dashed curves are tallies of an even number. The vertical hairlines show the parameter values held constant across panels within the same cluster.

doi:10.1371/journal.pone.0136088.g002

0.6), higher false-positive rate ($\alpha = 0.1$), and publication bias resulting in low communication of replications and negative findings. The range of base rates we show represents everything from genome wide association studies, on the low end ($b < 10^{-4}$), to predicting the winner of a presidential election, on the high end ($b = 0.5$). Every scientist will have a different opinion about which values represent realism. So in S1 File, we provide a Mathematica notebook for reproducing and altering these plots, so the reader can explore alternative scenarios of interest. But keep in mind that unrealistic scenarios are just as important for comprehending system dynamics.

First, notice that at tally $s = 1$ very many research findings are false. In the top cluster, the base rate must get quite high before a majority of hypotheses with tally $s = 1$ are true. In the bottom cluster, only the highest displayed base rates are sufficient. This dynamically replicates Ioannidis' direct calculation [1], even in the absence of bias and multiple testing. Many initially published findings are false, unless the base rate is high, and without any invocation of fraud or researcher bias.

Second, notice that replication helps, but how much it helps varies greatly. In the top cluster, even one positive replication at $s = 2$ renders most hypotheses true, at a base rate of $b = 0.1$. At lower base rates, $s = 3$ or $s = 4$ is required to raise precision above one-half. In the bottom cluster, low power and high false-positive rate make replication quite inefficient. Even at high base rates, $s = 3$ is needed. At low base rates, $s = 5$ or more is required. In either cluster, achieving near-certainty that an hypothesis is true always requires replication, even with a base rate as high as $b = 0.1$. In general, the same factors that make initial findings unreliable also make replications less reliable.

Note also that the rate of replication, r in panel (b), has remarkably little impact. This is because replication impacts the rate at which hypotheses reach different tallies, but not so much the precision at each tally. Therefore at low replication rates, few hypotheses will ever attain $s = 5$, but those that do are almost certainly true. We expand on this point in the next section.

Third, communication of findings, panels (e-g), can both assist discovery or hinder it. Suppression of negative replications (e) reduces precision. But suppression of positive replications (f) and novel negative findings (g) either improves precision or has almost no impact on it. These aspects of the population dynamics are counter-intuitive, but quite general and revealing. The next section explains them.

Dynamics of communication

The "file drawer problem" [13] arises when the failure to publish negative findings distorts the estimated strength of an association. We consider a related phenomenon by asking how changes in the communication parameters $c_{N\text{-}}$, $c_{R\text{-}}$, and c_{R+} alter the precision, sensitivity, and specificity across tallies. In the process, we'll have opportunity to explain the joint dynamics of research quality and communication biases.

In this model, it is rarely best to communicate everything. In S1 Text, we prove for the case of small b (such that $b^2 \approx 0$) and small r ($r^3 \approx 0$) that $c_{N\text{-}} < 1$ will improve precision when $\alpha < \beta$ (usually satisfied), that $c_{R\text{-}} < 1$ improves precision when $\alpha > \frac{1}{2}$ (hopefully never satisfied), and that $c_{R+} < 1$ improves precision whenever $\beta - \alpha \leq \frac{1}{4}$ (often satisfied). So some suppression of novel negative findings ($c_{N\text{-}} < 1$) and positive replications ($c_{R+} < 1$) can improve the value of replication. At larger b and r, the conditions are more complicated, but the qualitative finding remains intact.

To grasp why suppressing findings might help us learn what is true, think of replication as *epistemological chromatography*. Chromatography is a set of techniques for separating

substances that are mixed together. For example, mixed plant pigments can be separated by painting the mixture onto the tip of a strip of filter paper and then soaking the tip in a solvent. Different pigments bind more or less strongly to the solvent or the paper. Therefore as the paper absorbs the solvent, different pigments travel at different speeds, eventually separating and appearing as differently colored bands on the paper. In the epistemological case, it is true and false hypotheses that are mixed. We wish to separate the true ones from the false. Replication applies a "solvent" that diffuses false hypotheses towards negative tallies and true hypotheses towards positive tallies. A true hypothesis diffuses upwards with probability $(1-\beta)c_{R+}$, while a false hypothesis diffuses downwards with probability $(1-\alpha)c_{R-}$. Thus the communication parameters adjust rates of diffusion. Just as manipulating rates of chemical diffusion can improve real chromatography, manipulating communication can improve epistemological chromatography.

In Fig 3, we turn on communication one parameter at a time, in order to explain the contribution of each mode of communication to the resulting population dynamics. All four panels (a, b, c, d) show steady state precision, sensitivity, and specificity and use $b = 0.001$, $r = 0.2$, $1-\beta = 0.8$, and $\alpha = 0.05$. These values are chosen for clarity of illustration. In S1 File, we provide a Mathematica notebook to construct plots for any parameters the reader chooses. Note that for sensitivity and specificity, probability above/below the highest/lowest tally displayed is added up on the highest/lowest tally, so that none of the probability mass is hidden.

In the first three panels (a, b, c), only positive initial findings are communicated, and all new hypotheses appear at tally $s = 1$. The mixture of hypotheses at this tally is heavily skewed towards false hypotheses, and so has a low precision. Replication may cause an hypothesis to diffuse in either direction, depending upon communication. In panel (a), negative findings are never communicated. But since true hypotheses diffuse up at a rate $1-\beta$ and false ones only at a rate $\alpha < 1-\beta$, truth is slowly separated from falsity. At tallies of 8 or more, nearly all hypotheses are true, as indicated by the precision. Note however that most true hypotheses that have been communicated at all exist at low tallies, as indicated by the sensitivity. With enough time and replication, every true hypothesis can be split from the false. This is unlike the case in panel (b), where only negative replications are communicated. The same dynamic works in reverse here, and replication creates a pure sample of false hypotheses at low tallies.

Combining both directions of diffusion is synergistic, as illustrated in panel (c). Now both positive and negative replications are communicated. The downward diffusion of false hypotheses makes the upward diffusion of true hypotheses more efficient. This effect arises because $1-\alpha > 1-\beta$. False hypotheses diffuse down faster than true hypotheses diffuse up. This purifies the source mixture at $s = 1$, allowing for precision to approach high values at much smaller tallies than in the absence of either diffusion process. In this example, hypotheses with tallies of $s = 3$ and greater are true more than 80% of the time, and the sensitivity indicates that more than half of all published true hypotheses have a tally of 3 or more. Keep in mind that this 80% is equally interpretable as a probability that applies to a unique hypothesis. So it provides epistemic value, independent of the frequency interpretation.

Diffusion in both directions is enhanced by suppressing some positive replications. The dashed curves in panel (c) provide a comparison when only 20% of positive replications are communicated. Precision is substantially higher in this case, but at the cost of reduced sensitivity at high tallies. This effect arises from the same dynamic as before: by setting $c_{R+} < 1$, we have effectively slowed all upward diffusion. This allows rapid downward diffusion from negative replications to further clean the source mixture, but at the cost of diffusing more true hypotheses towards negative tallies. This dynamic is beneficial when base rate is especially low. So we achieve a very clean sample of truth at smaller positive tallies in this scenario, but at the

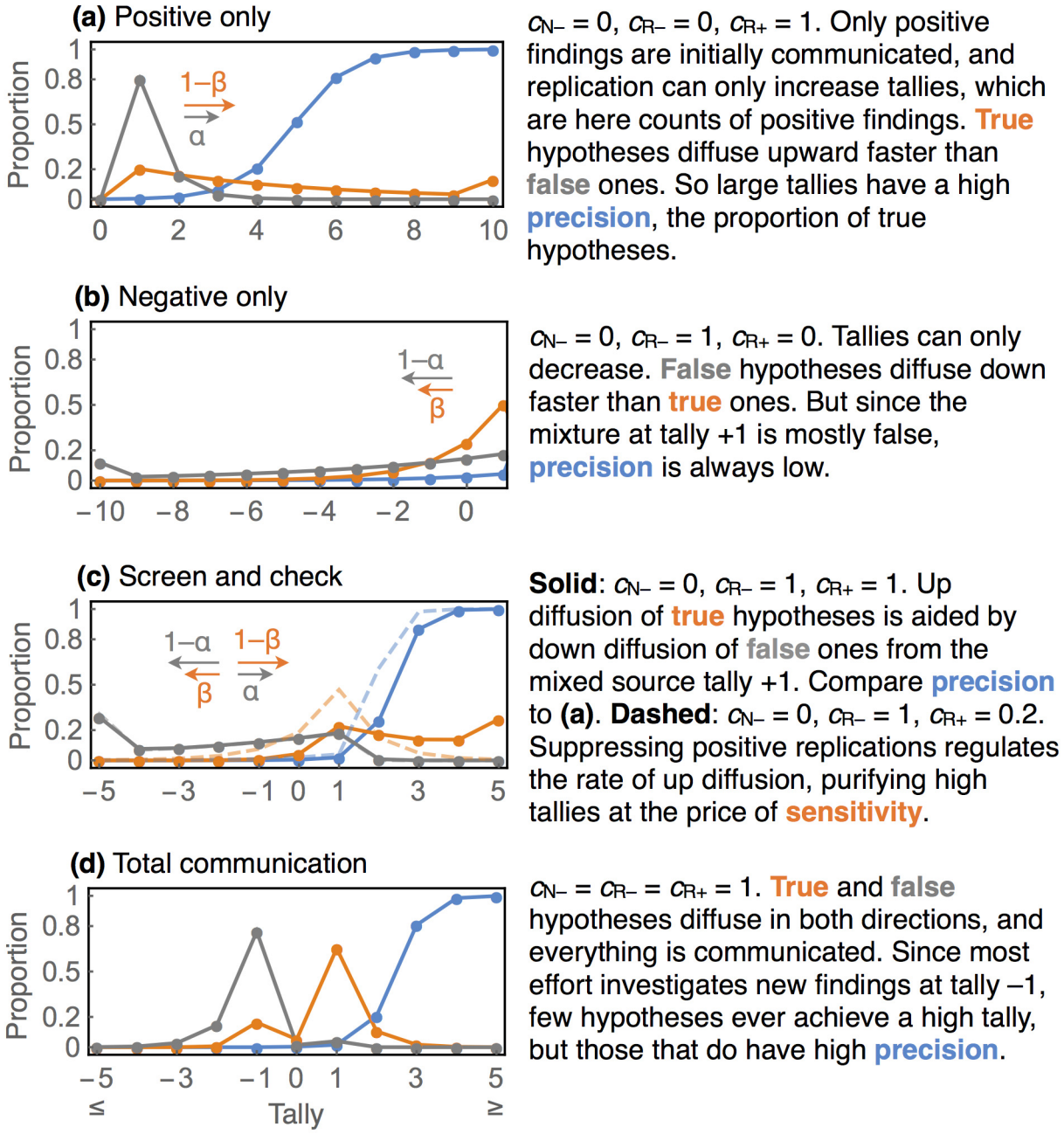

Fig 3. Replication and communication as epistemological chromatography. Precision is indicated in blue, sensitivity in orange, and specificity in gray.

doi:10.1371/journal.pone.0136088.g003

price of finding fewer true hypotheses in total. Whether this is an improvement depends upon context, an issue we take up in the discussion.

Finally, full communication is illustrated in panel (d). High precision is achieved at high tallies, but few hypotheses reside at those tallies. This inefficiency arises from the unbiased allocation of replication effort. When all initial findings are communicated, replication effort is overwhelmed by following up on initial negative findings, the spike in specificity seen at tally *s*

= −1. When the base rate is low, it can be better to screen for positive findings than to publish every negative finding. Note however that increasing precision, the proportion of hypotheses at a given tally that are true, is not necessarily the only objective. It does us little good if sensitivity is very low at all high tally values. We return to this point in a later section, when we consider differential power and false-positive rates between initial studies and replications.

Targeted replication

Replication in the preceding analysis is purely random: every communicated hypothesis has an equal chance of being the target of a replication effort. Targeting particular tally values, like $s = 1$, might be more efficient. Here, we demonstrate that the main effect of targeted replication is to improve sensitivity, the proportion of true hypotheses at positive tallies. It has little effect on precision, the proportion of hypotheses at positive tallies that are true.

To modify the population dynamics to allow targeted replication effort, assume that a proportion r_T of all replication attempts target a chosen list of tally values, selecting an hypothesis randomly from all hypotheses within the list. For example, this list might consist of all previously communicated hypotheses with a positive tally of three or less, so that researchers concentrate their replication efforts on hypotheses thought to be true but with relatively high uncertainty. The rest of the time, $1 - r_T$, replication effort remains unbiased.

Fig 4 shows the resulting modification of the dynamics. The dashed curves in these plots show the steady-state dynamics in the absence of targeting. The shaded pink regions show the range of tally values included in the target. In each case, targeting improves sensitivity at higher positive tallies. Thus it helps to diffuse true hypotheses towards tallies with very high precision. But there is very little effect on precision itself. Targeting helps because it directs effort towards tallies that may not have a high density of hypotheses. When replication effort is unbiased, most effort is directed to tallies where the bulk of hypotheses reside. Therefore when the target range includes a wide range, as in panel (c), it becomes relatively ineffective.

Why doesn't targeting improve the proportion of hypotheses that are true at higher tallies? Targeting serves mainly to speed up diffusion, without altering the *relative* rates at which true and false hypotheses diffuse. Changes in communication rates, in contrast, do alter the differential rates of diffusion, and so may dramatically alter precision, as seen in the previous section.

Differential power and false-positives

So far, we have assumed that power $1 - \beta$ and false-positive rate α are the same in initial studies and replications. Differences between initial studies and replications have been at the center of concerns about replication [9]. Here we analyze a version of our model in which we allow the power and false-positive rate to vary. Let $1 - \beta_R$ and α_R be the power and false-positive rate, respectively, for replications. What effects do both higher-powered replication and lower-powered replication have on dynamics?

In Fig 5, we present two extreme, illustrative scenarios. Both scenarios use $b = 0.001$, $c_{N-} = 0$, $c_{R-} = c_{R+} = 1$, $r = 0.2$, and $r_T = 0$ unless noted otherwise. The first is a "low/high" scenario in which initial findings are produced by studies with $1 - \beta = 0.6$ and $\alpha = 0.2$, but replications have conventional $1 - \beta_R = 0.8$ and $\alpha_R = 0.05$. This scenario reflects a context in which initial studies use small samples and suffer from motivated data-snooping or data-contingent analysis that elevates false-positives [22, 26]. This scenario is shown in panel (a). The second scenario is a "high/low" scenario, with $1 - \beta = 0.8$, $\alpha = 0.05$, $1 - \beta_R = 0.5$, $\alpha_R = 0.05$. This scenario reflects a context in which replications are prone to error, because a true effect requires skill to produce [9]. This scenario is shown in panel (b).

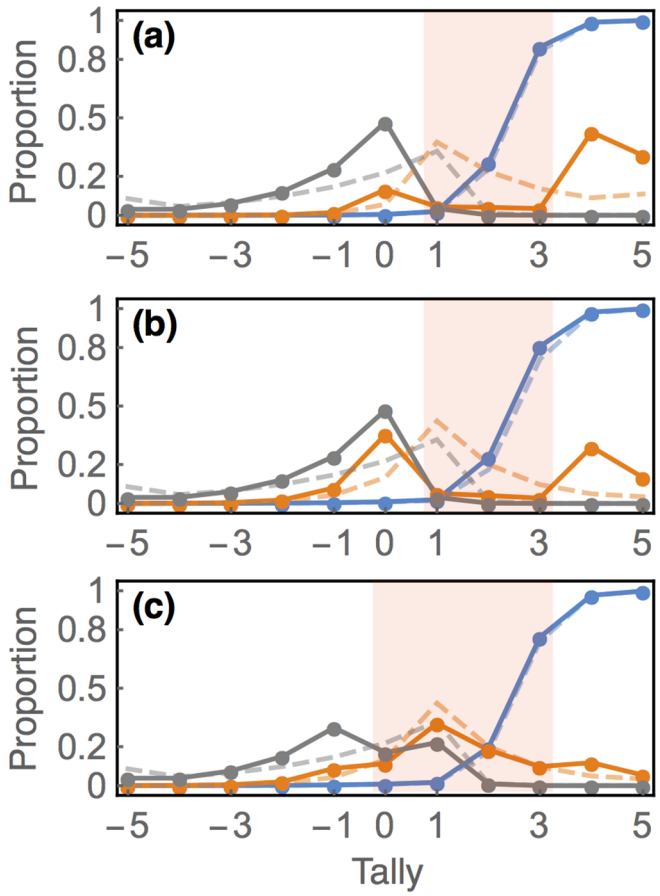

Fig 4. Targeted replication effort. In all three plots, tallies marked for targeted replication are shown by the shaded region. Precision is indicated in blue, sensitivity in orange, and specificity in gray. Baseline parameters set to $b = 0.001$, $\alpha = 0.05$, $r = 0.1$, $r_T = 0.5$, $c_{N-} = 0$, $c_{R-} = c_{R+} = 1$. Dashed curves display steady-state without targeted replication, $r_T = 0$. (a) High power setting, $1 - \beta = 0.8$. (b) Low power setting, $1 - \beta = 0.6$. (c) Low power, $1 - \beta = 0.6$, and including tally $s = 0$ in the target.

doi:10.1371/journal.pone.0136088.g004

Comparing the two, notice that low/high is more damaging overall, as the elevated false-positives cascade through the population during diffusion of hypotheses to higher tallies. Thus it takes more replication in (a) to achieve the same precision as in the high/low scenario (b). Even with only 50% power in (b), replication successfully separates true hypotheses from false ones. Unfortunately, it also diffuses many true hypotheses towards negative tallies. The high precision at positive tallies is a result of a false hypothesis' relative inability to attain a positive replication, not a result of a true hypothesis' ability to avoid a negative replication.

In the last two panels, (c) and (d), we show how these scenarios change when negative replications are suppressed, $c_{R-} = 0.1$. The situation generally worsens in both cases, but failure to communicate negative replications does prevent true hypotheses from attaining negative tallies, in the case in which replication power is low, (d).

Overall, replications continue to have value, even when they are more prone to error than original studies. As long as true hypotheses are more likely to diffuse upwards than downwards, replication aids discovery.

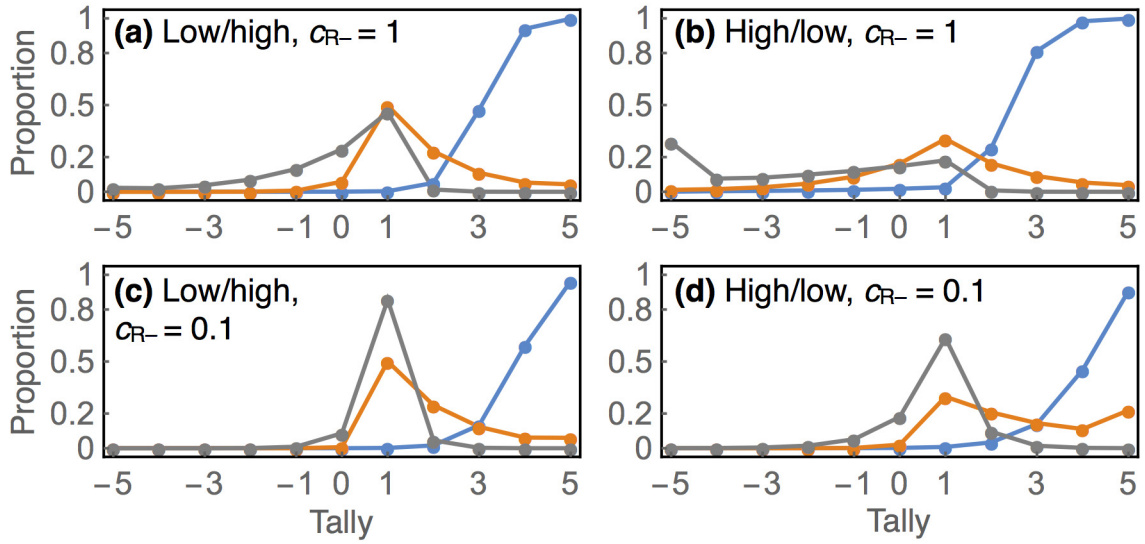

Fig 5. Differential power and replication dynamics. Precision is indicated in blue, sensitivity in orange, and specificity in gray. (a) Low power initial studies ($1 - \beta = 0.6$, $\alpha = 0.2$) but high power replications ($1 - \beta_R = 0.8$, $\alpha_R = 0.05$). (b) High power initial studies ($1 - \beta = 0.8$, $\alpha = 0.05$) but low power replications ($1 - \beta_R = 0.5$, $\alpha_R = 0.05$). (c) and (d) as in (a) and (b), respectively, but only 10% of negative replications are communicated.

doi:10.1371/journal.pone.0136088.g005

Discussion

Ours is the first analytical model of the joint population dynamics of scientific hypothesis generation, communication, and replication. Such a model is necessary to illuminate debates about scientific practice, because until researchers report the results of every study, empirical estimates of base rate are not possible. And without consideration of population dynamics, any discussion of the value of research findings remains at least partly naïve, because it is notoriously difficult to reason verbally about complex systems. Our model produces a number of valuable counter-intuitive results. But even when its results are intuitive, some model like ours is needed to demonstrate their logic. It is not enough to merely hold the correct belief; we must also justify that belief.

This model is not a definitive representation of the scientific process, nor does it aim to be. It omits many relevant factors, such as investigator bias and disagreements about the interpretation of evidence. These omissions allow the model to address focused questions about the evidential value of research as it emerges from the joint dynamics of hypothesis generation, replication, and communication. Models that account for more and different factors must also include variants of these complex dynamics, so our model is a necessary and useful first step.

Our analysis re-emphasizes what every textbook says: replication is an essential aspect of scientific discovery. However, it also quantifies its impact and emphasizes that replication itself can be unreliable—the factors that make initial findings unreliable also make replication less reliable. When base rate is low, power is low, or false positives common, then many successful replications will be needed to attain confidence in an hypothesis. This is especially true when negative replications are difficult to publish.

We find that low base rate and high false positive rate are the most important threats to the effectiveness of research, replicated or not. This re-emphasizes the importance of quality theorizing, in order to improve base rate. While it is appealing to think that science works

regardless of where hypotheses come from, undisciplined hypothesis generation reduces base rate and makes initial findings mostly false. Then large amounts of replication will be needed to uncover the truth. In fields such as physics and evolutionary biology, a great deal can be and is done to vet theory in the realm of pure thought, using mathematics and simulation. But in fields such as social psychology, theory development is rarely formalized [27].

The results also re-emphasize the value of efforts to suppress false positive findings, such as pre-registered data analysis plans. It is important to recognize that any single scientific hypothesis may correspond to many different statistical hypotheses. If a statistical hypothesis can be chosen after seeing the data, reasonable scientific hypotheses can become unreasonably flexible [28]. And many data-contingent transformations and modeling choices that increase power, conditional on an hypothesis being true, will also increase false-positives, conditional on the hypothesis being false. For example, dropping outliers may well aid discovery, if the hypothesis is true. But it may also dramatically inflate false-positives, if the hypothesis is not true [29].

Our model immediately informs debates over the meaning of failed replications. For example, some have suggested that positive replications have more worth than negative replications [12], or even that failed replications "cannot contribute to a cumulative understanding of scientific phenomena" [30]. We find the opposite: communicating a failure to replicate is typically more informative than communicating a successful replication. This remains true even when replication attempts have lower power than original studies. However, a single failure to replicate is entirely consistent with a true hypothesis in many scenarios. So both positive and negative replications may be regarded with skepticism. But neither is without value. Of course our model is merely a model. But unlike the verbal arguments we cite, it is at least clear in its assumptions, and its logic can be verified.

Our model also sheds light on proposals for improving the reliability of research. For example, many have called for pre-registration and review with a commitment from journals to publish research results, positive or negative, in order to reduce under-reporting of negative findings [31]. Our analysis suggests that these proposals should distinguish between new hypotheses and replication attempts. If indeed many new hypotheses are false in many fields, a pre-registration process would merely fill journal pages with null findings, doing great harm by crowding out candidate hypotheses that have passed an initial screening. In our model, there is little harm in ignoring novel negative findings, because they add very little information. Indeed, Fig 2 illustrates that the effect of ignoring novel negative results on precision is negligible. In contrast, a negative replication may add a lot of information. We suspect however that our model exaggerates this effect, because the model ignores the wasted effort arising from different researchers repeating an investigation in ignorance of one another's negative findings. And there are certainly fields in which full publication may be the best policy, such as when false-positive rates are low or when the total number of testable hypotheses is very small. Nevertheless, the qualitative difference in information value between novel and follow-up negative findings will remain as long as the base rate in the published literature is higher than it is in novel investigations.

The model stimulates empirical investigation by clarifying which factors must be estimated in order to gauge the evidential value of research, as well as being readily translatable into a statistical framework, due to its analytical specification. Our model provides an implicit 'null model' of research: setting $b = 0$ provides a null distribution of novel findings and lifespans of hypotheses. Null models are deliberately unrealistic and usually *a priori* false, but have nevertheless played an important role in science [20].

There are additional factors to address in future work. Our model ignores researcher bias, multiple testing, and data snooping, each of which deflates base rate or inflates false-positive rate. Our analysis is framed in a standard, but unsatisfying, "true" and "false" classification,

rather than considering practical significance and effect size estimation [26]. Our model can be directly generalized to consider variation in effect size instead of true and false hypotheses. We explain this generalization in S1 Text. However, our model does not directly address causal inference nor point estimation.

Incentives also matter. A dynamic analysis of strategic behavior under different incentive structures would aid policy analysis [18]. As Karl Popper argued, science does not work because scientists are selfless and unbiased people. Rather it works because its institutions channel our bias into the production of public goods [32]. In particular, we worry that a research environment that lacks replication may actually select for statistical practices that inflate false-positives, as labs with such practices can more readily publish findings and place students in new positions, all while outrunning the truth.

Replication may offer other benefits that are not accounted for in our model. A failed replication may be valuable because it inspires a new hypothesis in order to explain variation in findings. When findings do not generalize across samples, this creates an opportunity to explain the variation [33, 34]. In our view, the goal of replication is not merely to find the same result, but also to discover how a result arises and how it is likely to vary in realistic, non-laboratory, contexts.

Despite these shortcomings, our model provides specific quantitative evaluations of many verbal arguments, as well as drawing attention to the population dynamics of scientific knowledge. Science is a subtle project. Understanding it demands the same rigor that we apply to projects within science itself.

Supporting Information

S1 Text. Mathematical details, including derivation of steady-state solutions.
(PDF)

S1 File. Mathematica notebook, with model equations and figures.
(NB)

Acknowledgments

We thank our colleagues at UC Davis, audiences at UC Berkeley's Institute for Data Science and the SPSP pre-conference on Dynamical Systems and Computational Modeling, and Titus von der Malsburg for helpful comments on a pre-print version of this paper.

Author Contributions

Conceived and designed the experiments: RM PS. Performed the experiments: RM PS. Analyzed the data: RM PS. Contributed reagents/materials/analysis tools: RM PS. Wrote the paper: RM PS.

References

1. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005 Aug; 2(8):e124. Available from: http://dx.doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
2. Makel MC, Plucker JA, Hegarty B. Replications in Psychology Research How Often Do They Really Occur? Perspectives on Psychological Science. 2012; 7(6):537–542. Available from: http://pps.sagepub.com/content/7/6/537 doi: 10.1177/1745691612460688 PMID: 26168110
3. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. Science. 2014; 345:1502–1505. Available from: http://www.sciencemag.org/content/345/6203/1502 doi: 10.1126/science.1255484 PMID: 25170047

4. Schmidt S. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology. 2009; 13(2):90–100. doi: 10.1037/a0015108
5. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483:531–533. Available from: http://www.nature.com/nature/journal/v483/n7391/full/483531a.html doi: 10.1038/483531a PMID: 22460880
6. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011; 10(9):712–712. Available from: http://www.nature.com/nrd/journal/v10/n9/full/nrd3439-c1.html doi: 10.1038/nrd3439-c1 PMID: 21892149
7. Sullivan PF. Spurious Genetic Associations. Biological Psychiatry. 2007 May; 61(10):1121–1126. Available from: http://www.sciencedirect.com/science/article/pii/S0006322306014703 doi: 10.1016/j.biopsych.2006.11.010 PMID: 17346679
8. Fontani M, Costa M, Orna MV. The Lost Elements: The Periodic Table's Shadow Side. Oxford University Press; 2014.
9. Bissell M. Reproducibility: The risks of the replication drive. Nature. 2013; 503:333–334. Available from: http://www.nature.com/news/reproducibility-the-risks-of-the-replication-drive-1.14184 doi: 10.1038/503333a PMID: 24273798
10. Bohannon J. Replication effort provokes praise—and 'bullying' charges. Science. 2014; 344:788–789. Available from: http://www.sciencemag.org/content/344/6186/788 doi: 10.1126/science.344.6186.788 PMID: 24855232
11. Kahneman D. A new etiquette for replication. Social Psychology. 2014; 45:310–311.
12. Schnall S. Clean data: Statistical artefacts wash out replication efforts. Social Psychology. 2014; 45 (4):315–320. Available from: http://www.psycontent.com/content/k5257g3605571477/ doi: 10.1027/1864-9335/a000204
13. Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin. 1979; 86 (3):638–641. doi: 10.1037/0033-2909.86.3.638
14. Hull DL. Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science. Chicago, IL: University of Chicago Press; 1988.
15. O'Rourke K, Detsky AS. Meta-analysis in medical research: Strong encouragement for higher quality in individual research efforts. Journal of Clinical Epidemiology. 1989; 42(10):1021–1024. doi: 10.1016/0895-4356(89)90168-6 PMID: 2809651
16. Campbell DT. Toward an epistemologically-relevant sociology of science. Science, Technology, & Human Values. 1985; 10(1):38–48.
17. Popper K. Conjectures and Refutations: The Growth of Scientific Knowledge. New York: Routledge; 1963.
18. Kitcher P. Reviving the Sociology of Science. Philosophy of Science. 2000; 67:S33–S44. doi: 10.1086/392807
19. Levins R. The Strategy of Model Building in Population Biology. American Scientist. 1966; 54.
20. Wimsatt WC. False Models as means to Truer Theories. In: Nitecki M, Hoffman A, editors. Neutral Models in Biology. London: Oxford University Press; 1987. p. 23–55.
21. Munroe R. "Significant": http://xkcd.com/882/; 2014. Available from: http://xkcd.com/882/ [cited 2014].
22. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science. 2011; 22 (11):1359–1366. Available from: http://pss.sagepub.com/content/22/11/1359 doi: 10.1177/0956797611417632 PMID: 22006061
23. Cox RT. Probability, Frequency and Reasonable Expectation. American Journal of Physics. 1946; 14:1–10. doi: 10.1119/1.1990764
24. Sedlemeier P, Gigerenzer G. Do studies of statistical power have an effect on the power of studies? Psychological Bulletin. 1989; 105(2):309–316.
25. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14(5):365–376. Available from: http://www.nature.com/nrn/journal/v14/n5/full/nrn3475.html doi: 10.1038/nrn3475 PMID: 23571845
26. Gelman A, Loken E. Ethics and Statistics: The AAA Tranche of Subprime Science. CHANCE. 2014; 27 (1):51–56. Available from: http://amstat.tandfonline.com/doi/abs/10.1080/09332480.2014.890872 doi: 10.1080/09332480.2014.890872
27. Smaldino PE, Calanchini J, Pickett CL. Theory development with agent-based models. Organizational Psychology Review. 2015;in press.

28. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University; 2013.

29. Bakker M, Wicherts JM. Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. Psychological Methods. 2014; 19:409–427. doi: 10.1037/met0000014 PMID: 24773354

30. Mitchell J. On the emptiness of failed replications; 2014. Available from: http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm

31. American Political Science Association Task Force on Public Engagement. Increasing the credibility of political science research: A proposal for journal reforms; 2014.

32. Popper K. The Myth of the Framework: In Defence of Science and Rationality. Routledge; 1996.

33. Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, Bolyanatz A, et al. Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. Science. 2010; 327:1480–1484. Available from: /rmpubs/henrichetalfairnessmarketsreligiongroupsizeScience2010.pdf doi: 10.1126/science.1182238 PMID: 20299588

34. Scott IM, Clark AP, Josephson SC, Boyette AH, Cuthill IC, Fried RL, et al. Human preferences for sexually dimorphic faces may be evolutionarily novel. Proceedings of the National Academy of Sciences. 2014; 111(40):14388–14393. Available from: http://www.pnas.org/content/111/40/14388.abstract doi: 10.1073/pnas.1409643111