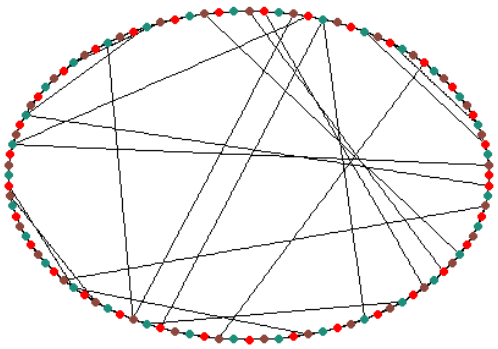
The Truth Hurts

Investigating replication, diversity and consensus in science
using agent-based modelling



Victor Møller Poulsen (au591295@post.au.dk)
Undergraduate at Cognitive Science, Aarhus University

Mikkel Werling (au591251@post.au.dk)
Undergraduate at Cognitive Science, Aarhus University

AARHUS
UNIVERSITY

Abstract

This paper expands upon a recent agent-based model (ABM) of scientific discovery published by Devezer et. al (2019). They conclude that replicability is not a good measure of other desirable properties of science, that innovation speeds up discovery of truth and that diversity promotes a healthy science. We follow Devezer et. al (2019) in modelling science generally, but some specifications in our model are aimed at capturing the state of affairs in the social sciences, and more specifically in psychology. The focus of this paper is on replicability, diversity and consensus in scientific communities. Our main addition to their model lies in situating agents in a network structure. Our ABM replicates their findings regarding the disconnect between desirable properties of science and replication rate. It also replicates the connection between innovation and early discovery of truth. However, the benefit of diversity is only partly generated by our model. We investigate additional questions regarding consensus in science and find that scientific consensus surrounding truth is the exception, not the rule. Our system generates replication rates which match the rates seen in psychology and other social sciences. Furthermore, we find that although replicability is not a good proxy for truth, true results are more likely to replicate.

Keywords: Agent-based modelling, meta-science, replication, diversity, consensus

Table of contents

Abstract 1
1. Introduction 3
2. The original framework (Mikkel, Victor) 4
3. Key issues (Victor, Mikkel) 5
4. The network solution (Mikkel, Victor) 6
4.1. The issue of the global model (Victor, Mikkel) 6
4.2. The issue of replication (Mikkel, Victor) 7
5. Why ABM? (Victor, Mikkel) 7
5.1. How ABMs work (Mikkel, Victor) 8
5.2. How our ABM works (Victor, Mikkel) 8
6. Hypotheses (Mikkel, Victor) 9
7. Implementation (Victor, Mikkel) 10
7.1. Alterations (Mikkel, Victor) 11
7.1.1. Types and populations 11
7.1.2. Communication 11
7.1.3 Time-steps 12
7.2. Constants (Victor, Mikkel) 12
7.2.1. Models 12
7.2.2. Model comparison 13
7.2.3. Networks 13
7.3. Manipulations (Mikkel, Victor) 15
7.3.1. Sample size 15
7.3.2. Sigma 15
7.3.3. Replications 15
7.4. Probabilistic and random events (Victor, Mikkel) 16
7.4. Outcome measures (Mikkel, Victor) 16
8. Results (Mikkel, Victor) 17
8.1. The benefits of diversity (Victor, Mikkel) 18
8.2 Innovation drives the speed of progress (Mikkel, Victor) 19
8.3 Replication rate & desirable scientific properties (Mikkel, Victor) 19
8.4 Low replication rate (Victor, Mikkel) 20
8.5. Reaching consensus (Mikkel, Victor) 22
9. Limitations (Victor, Mikkel) 23
9.1. Microspecifications (Victor, Mikkel) 23
9.2. Information cascades (Mikkel, Victor) 24
10. Conclusion 25
12. Appendix 26
13. Literature 29

1. Introduction

Although the word "truth" does not figure prominently in the debate surrounding replications, it is clear that replications are used to indicate whether findings are true (Goodman, Fanelli, & Ioannidis, 2016). Independent confirmations of scientific findings are at least taken to lend credibility to scientific claims. Similarly, when a scientific community consistently fails to obtain such confirmations our trust in its literature is eroded. When this happens, the very notion that we are making scientific progress is called into question. In this paper we will refer to confirmations of scientific findings as "replications", though some have used “reproducibility” to mean the same thing (Goodman et al., 2016).

In psychology the debate over replicability has been at the centre of attention since a massive collaboration failed to replicate the results of 64% of targeted findings from top psychology journals (A. Aarts et al., 2015). However, the worry that scientific literatures are brimming with false findings is not confined to psychology. Ioannidis (2005) focuses on biomedical research and concludes that most findings are likely to be false. He argues that this is largely due to low power resulting from small sample sizes, small effect sizes, as well as the prevalence of null hypothesis significance testing. Other frequently proposed drivers of high false positive rates include incentive structures, researcher bias and questionable research practices (QRPs) (A. Aarts et al., 2015; Ioannidis, 2005; Smaldino Paul E. & McElreath Richard, 2016).

In this paper we ignore issues of incentive structures and scientific misconduct. We are interested in whether replication rate is a good measure of truth when no such issues exist. Thus, we model agents who do not conduct questionable research and who operate in a world without incentive structures. We do not regard replication rate as an intrinsically desirable property of science, but as desirable only insofar as it correlates with actually desirable properties of science.

In addition to exploring the association between replication rate and desirable scientific properties, we explore whether diversity promotes desirable scientific outcomes. In his book "Diversity and Complexity", Scott. E. Page (2010) makes a compelling case for the many benefits of diversity in securing robustness and performance of complex systems. Here we focus on two of them: averaging and diminishing returns to type. Averaging points to the fact that a diverse population consisting of agents who individually perform well given different circumstances will never perform terribly on average given any circumstance. The performance of the system will be an average of high-performers, low-performers and everything in between. In that sense, the system will have robustness against worst outcomes.

Diminishing returns to type is something that we are all familiar with from various settings. The first slice of cake has a higher utility value than the second. And the second slice has a higher utility than the third. The same is true in science, where the first mathematician you add to your research team will add more value than the second and third. Thus, a research team with a diverse population will be likely to outperform any non-diverse population. A research team of one philosopher, one mathematician and

one psychologist can outperform the best performing team consisting of either three philosophers, three mathematicians or three psychologists. In this sense diversity can offer not just robustness but a net benefit for the diverse population.

2. The original framework (Mikkel, Victor)

One of the most recent attempts at modelling scientific discovery was conducted by Devezer et. al (2019). Their model provides a solid framework for investigating replication and epistemic diversity. We introduce their system and main findings before presenting our additions. Devezer et. al (2019) investigate scientific discovery through an agent-based model (ABM) where different agents attempt to find a true model. They refer to this as a model-centric framework of science. The process of their ABM can be described in five sequential steps:

1. Each time-step in their model starts with the sampling of an agent from a population of agents. This agent is referred to as being the agent on turn. Each agent represents a scientist with a specific research strategy.

2. The scientist samples a model based on their research strategy and what the current scientific consensus is. This model is called the "proposed model". The scientist can sample from 14 different models out of which one is the true model.

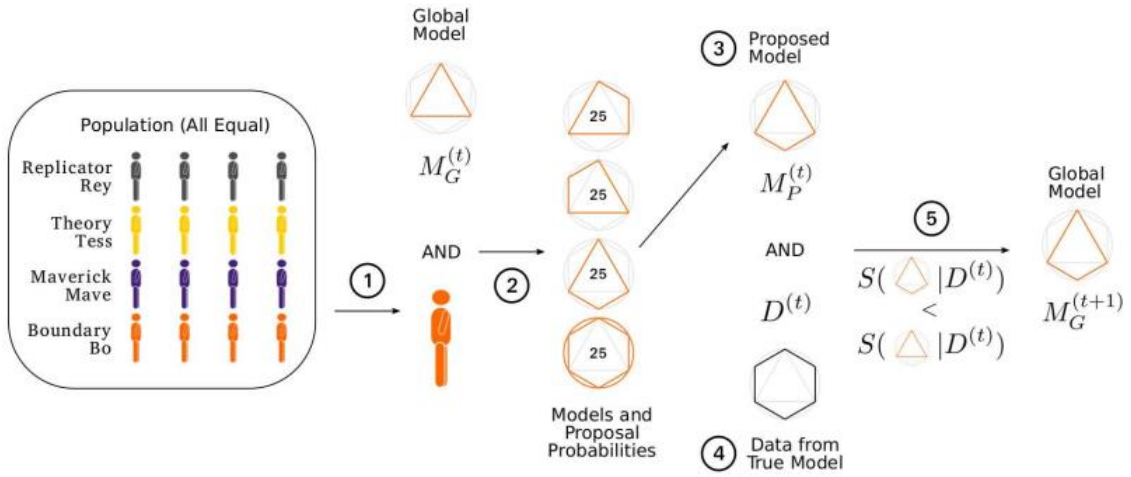
3. Next, data is generated from the true model. The true model is thus the model which accurately describes the process which generates the data that the scientists are sampling from.

4. Now the "proposed model" is compared to the scientific consensus - the "global model". This is done using information criterion, to determine which of the two models is most likely to be true.

5. If the proposed model has the lowest information criterion score then it will become the new scientific consensus - the new "global model". If the proposed model has the highest information criterion score then nothing changes and the next agent is sampled.

The described 5-step process represents one time-step in their model. Devezer et. al (2019) run 11.000 time-steps per replica. The process is schematically represented below (Fig. 1).

Figure 1



Taken from Devezer et. al (2019). The picture explains the process of change from the global model to the proposed model. (1) A scientist is sampled from a population of scientists. (2) Based on which scientist type is sampled and what the current global model is, (3) a new model is proposed. (4) The proposed model is compared to the global model using information criterion. (5) If the proposed model has a lower information criterion score than the global model, the proposed model becomes the new global model.

3. Key issues (Victor, Mikkel)

As is the case with all models, the model presented in Devezer et. al (2019) does some violence to reality. A model is not supposed to be the same as reality but a good model captures the most salient properties while exclude all irrelevant factors (Smaldino, 2017). We identify two modelling choices in Devezer et. al (2019) which inhibit us from answering our key questions satisfactorily. The first problematic design choice is the construct of a *global model*. This modelling choice comes with several unrealistic assumptions about how science operates.

Firstly, the global model assumes perfect sharing of information between the different scientists in the community. In the real world, no scientist can read about every new finding or collaborate with all colleagues in her field.

Secondly, the global model assumes that communities of scientists adhere to the same theory of the phenomenon at hand. Although the natural sciences show a high degree of consensus, the social sciences do not show this pattern. Clusters of scientists which support different schools of thought within particular domains are the norm (Collins, 1994). Besides being unrealistic assumptions, the construct of a

global model also results in an inability to track incremental progress towards the truth, which is something we want to track.

The second problematic design choice concerns which studies are being replicated. In the original model, replications happen when their replication agent (a research strategy) is sampled and conducts a replication of the most recently conducted study with new data.

Because the replication agent is sampled randomly, this design choice assumes that scientists choose findings to replicate randomly - and on average - equally often. There are two ways in which this assumption is unrealistic. First, many of the replication studies conducted in the model will compare models where one is clearly superior to the other. Second of all, many of the replication studies conducted in the model will be done on previous studies where the global model did not change (I.e., null findings). In the real world scientists do not spend resources testing obviously good models against obviously bad ones, and null findings are unlikely to be published (much less replicated). In the real world replications are done mainly on novel research findings from the published literature (A. Aarts et al., 2015). We believe that this modelling choice should result in an inflated replication rate compared to what we observe in the real world.

4. The network solution (Mikkel, Victor)

We sought to alleviate these issues by implementing the scientific community in a network structure. One of the network structures that has increased in popularity for the modelling of social structures is the small-world network. Particularly, it has also previously been used to model scientific communities (Ebadi & Schiffauerova, 2015). The small-world network can be viewed as a network, where agents are mostly connected with their neighbours, but also have a probability to add a connection to a random agent in the network in exchange for a connection to one of their neighbours (Watts & Strogatz, 1998). The network structure allows us to model each scientist as an individual entity within the scientific community. The specifics of this network will be explored in greater detail in its own section.

4.1. The issue of the global model (Victor, Mikkel)

Situating our agents in a network permits us to exchange the construct of a global model with multiple *local models.* Where the global model represents the current belief of the global community, the local model represents the local belief of each individual scientist. Proposed models are thus compared to the scientists' own beliefs. When the proposed model obtains a lower information criterion score than their

local model, the local model is changed to the proposed model. The substantial change is that scientific findings alter local beliefs instead of changing the belief of the entire community.

The change from a global to many local models allows us to circumvent perfect information sharing. Agents only share their beliefs with the agents that they are connected to in the network. This constraints communication to local clusters of the scientific community.

The network implementation also allows us to make more realistic assumptions regarding consensus in science. As argued above, consensus is rare in the social sciences, and we are thus interested in tracking proportions of agents who believe in the true model. As each agent has its own local model we can calculate the proportion of agents for whom the true model is also their local model at each time-step. By tracking this rate over time, we can investigate scientific progress as an incremental process. This information was inaccessible under the construct of a global model.

4.2. The issue of replication (Mikkel, Victor)

Our way of selecting studies for replication differs from the mechanism used in Devezer et. al (2019). In our network a replication study happens under very specific circumstances. Firstly, an agent has to change her local model. This inspires her neighbours to test this new model against their local model. Because replication is conditional on a neighbour making a research finding, we do not have any replicator agent. Instead all other agent types can become replicators when they get inspired by their neighbours.

This design assumes that replications happen when someone you are connected to finds something new. We believe that this way of modelling replication should make our replication rates lower than those found in Devezer et. al (2019) and thus more reflective of typical replication rates in the social sciences. This is because there won't be any replications of null-findings and fewer replications in which one of the tested models is obviously inferior.

5. Why ABM? (Victor, Mikkel)

Now we turn to the methodology used to implement the framework which we have loosely described. Agent-based modelling (ABM) is a method in which one simulates actors and measure system-level effects of their interactions (Crooks & Heppenstall, 2012). Agent-based modelling is typically used to investigate complex systems. Scott. E. Page defines a complex system as a system consisting of "*di-*

verse, rule following *entities* whose behaviours are *interdependen*t. Those entities interact over a *contact structure* or *network*. In addition, the entities often *adapt*." (Page, 2010, p. 16; his) Communities - such as scientific communities - are complex systems.

5.1. How ABMs work (Mikkel, Victor)

The way that the modeller approaches the study of complex systems is to identify a target macrostructure and generate it from a microspecification. The macrostructure refers to target system-level effects. For instance, it is the case that in many scientific communities we see (1) a lack of consensus and (2) low replication rates of published findings. These are system-level effects that we expect our model to generate. The microspecification refers to the behavioral rules that entities - or agents - follow in the model.

Agent-based models differ from other scientific work in that they are by definition generative. It is not enough to explain why macrostructures stay the way they are once they have converged to some state of stability. You also need to show how the macro-structure could have developed in the first place. Epstein (1999) refer to this as growing the macroscopic structure from the "bottom up".

Although it represents a distinct methodology, ABM research intersect with other empirical work. To device an agent-based model one needs a realistic microspecification. The macrostructure generated by the model is contingent on a realistic microspecification. The input should be specified in cooperation with experts in the particular area which is being modelled (Epstein, 1999). Devezer et. al (2019) device agents which follow simple behaviours that they argue correspond roughly to classical research strategies. We will also refer to relevant literature when considering input specifications to our model in the coming section.

The output of agent-based models also interfaces with other empirical science. If an agent-based model successfully reproduces a target macrostructure then it shows that the microspecification is *sufficient* to generate the observed system-level properties. It is then a candidate explanation. However, there might be several ways of generating a given system-level property. If several mechanisms suffice then the most realistic microspecification should be given priority as the preferred explanation (Epstein, 1999).

5.2. How our ABM works (Victor, Mikkel)

Our model follows many typical features of agent-based models listed in Epstein (1999). Firstly, the agents in our population are heterogeneous. The heterogeneity comes from our different types of agents. Our agents also act autonomously in the sense that there is no central executive or "top-down" control. Both properties are also shared by the model in Devezer et. al (2019).

However, the model in Devezer et. al (2019) is a very limited agent-based model. In our model agents are represented in explicit space, the small-world network. In addition our agents are bounded. They are bounded in the sense that they can only interact with the agents they are connected to and in the sense that they do not have access to perfect information. In these respects our agent-based model is more complex than the model reported in Devezer et. al (2019).

There are two aspects in which our agent-based model could be made more dynamic. As hinted at in the definition of a complex system, dynamic systems often have agents who adapt (Page, 2010). Our agents do not adapt but stick to their rules of behaviour throughout the simulation. The microspecification is static. In addition, our network does not change throughout each iteration. The agents are placed in a small-world in which they have certain connections. They do not have the option of creating new connections or losing old ones. The connectivity pattern of the network is static. Adding either of these dynamical properties would make the model more ecologically valid, but we judge that these properties are not necessary for us to answer the key questions that we pose at this stage.

A last point about ABMs which is salient to our project is that they often take aim at generating phenomena qualitatively. Epstein calls these demonstrations "stylized facts" (Epstein, 1999). Our replication of Devezer et. al (2019) is qualitative in the sense that we have not designed our model so as to compare our quantitative outputs against theirs. We are interested in whether the tendencies that they report replicate in a network structure. Furthermore, we are interested in whether certain macrostructures of interest are generated by our model. As mentioned earlier these include (1) a lack of scientific consensus and (2) low replication rates. Only to a lesser degree are we interested in comparing whether our replication rates quantitatively match those of e.g., psychology. Although we believe that our ABM improves and extends the one in Devezer et. al (2019) it is still too crude to allow for meaningful quantitative comparisons.

6. Hypotheses (Mikkel, Victor)

The implementation of the network structure and our redefined mechanism for replication studies constitute our improvements to the framework of the original paper. We sought to make a conceptual replication of the three main findings of Devezer et. al (2019):

H_1: Epistemically diverse populations will never be the worst population on any measure of desirable scientific properties. These measures include:

- first passage time to true model.

- time spent at the true model (what we call “agents at the true model).
- stickiness of the true model.
- replication rate (which they classify as a desirable scientific property).

H_2: Innovative strategies will reach the true model at a faster pace than the other populations. The innovative strategy is represented by the agent type Mave, which will be introduced in the next section.

H_3: There is no correlation between replication rate (what they refer to as rate of reproducibility) and first passage time to true model, time spent at the true model and stickiness of the true model.

In addition we consider two additional questions. Both are related to macrostructures that we see in scientific fields such as psychology and that we expect will emerge from our microspecification.

H_4: We will see a lack of consensus among agents as to what the true model is. Generally, we expect that less than 50% of the scientific field will have the true model as their local model at any given time.

H_5: We expect to see a low replication rate. Devezer et. al (2019) report quite high replication rates (between 89.9 and 100%). We expect our replication rates to be lower, approaching the 36% replication rate found in the big psychology replication project mentioned earlier (A. Aarts et al., 2015).

7. Implementation (Victor, Mikkel)

The implementation of our ABM will be explained in terms of alterations of the original framework, the constants of the implementation and the manipulations of different variables.

Alterations: We altered the number of types of agents, how agents communicate with each other and the number of time-steps for replica.

Constants: We kept the number of models, the method of model comparison, and the generative process of the networks constant.

Manipulations: We manipulated the populations, the sigma level, the true model and the sample size.

7.1. Alterations (Mikkel, Victor)

7.1.1. Types and populations

To explain our alteration, we must first introduce the different types of scientists. As was alluded to earlier, each type of scientist represents different research strategies. The proposed model depends on the current local model of an agent and what type of scientist they are. This is since the different scientist types use their current beliefs to inform them about what kind of model they should propose. We operate with three different scientist types all based on Devezer et. al (2019):

1. *Tess, the theory tester:* Tess proposes a model which differs by one main effect from the current local model.

2. *Bo, the boundary tester:* Bo proposes a model which adds one or more interaction effects to the current local model.

3. *Mave, the maverick:* Mave ignores her local model, and proposes a random model. Mave's research strategy represents an innovative research strategy, as she ignores her previously held beliefs and samples the model space without bias.

In Devezer et. al (2019) *Rey, the replicator* is the fourth agent type. However, we have excluded this strategy based on the considerations of replication discussed earlier.

From these three scientist types, four different populations are made. Three of them are populations where one scientist type constitutes 98% of the population, and the final 2% is divided equally among the two other types of scientists. Note that since we have 100 agents in our networks, these proportions correspond to the actual number of agents in our network. The three populations described are homogeneous as they mainly consist of one agent type. They are referred to as being dominated by one particular research strategy. These populations are the *Tess*-dominant population, the *Bo*-dominant population and the *Mave*-dominant population. The final population is an *epistemically diverse* population, in which all scientist types are represented in equal proportion. In essence, population refers to what type of research strategy is dominant in the particular scientific community.

7.1.2. Communication

In order to constrain how agents communicate with each other, we change the five-step process described in the framework section (Fig. 1) with two considerable alterations. The first change is the previously mentioned alteration to the fifth step of the process, which replaces the global model with a local model. The second is an optional sixth step to the process, which is permitted by our network

structure. The process only enters the sixth step when a local model is changed. When this happens, all the scientists who are connected to the scientist on turn are activated. These scientists then compare their own local model to the newly proposed model. In effect they are doing a new study inspired by the finding of their colleague. If the proposed model outperforms their local model, they replace their local model with the proposed model. This happens inside the time step of the original agent on turn. Note that it is only the agents connected to the agent on turn who test this newly sampled model. As such our design does not allow for information to cascade through the system.

7.1.3 Time-steps

In the original framework one run of the 5-step process described in the framework section (Fig. 1) is called a time-step. They call a single run of 11.000 *time-steps* a replica. Each replica of their simulation equates to 11.000 studies being conducted by the scientists. In our implementation each time-step does not equate to a study. If the agent on turn changes her local model all other agents connected to her will also test this new model. They are performing a study just as much as the agent on turn. This inflates the number of studies being done at each time-step. We wanted to limit the number of studies per replica to 11.000, as conducting and planning studies is arguably the time-consuming aspect of science. We call a single run of 11.000 *studies* a replica. We still rely on time-steps for certain outcome measures. We based the length of each replica on amount of studies for two reasons:

1. Some populations change models more than others. The result of which is that some populations simply conduct more studies per time-step than the other populations. Limiting studies instead of time-steps effectively levels the playing field by ensuring that each replica will have the same amount of computations.

2. We wanted to make our results comparable with those of Devezer et. al (2019) and adding additional studies would complicate a comparison more than necessary.

7.2. Constants (Victor, Mikkel)

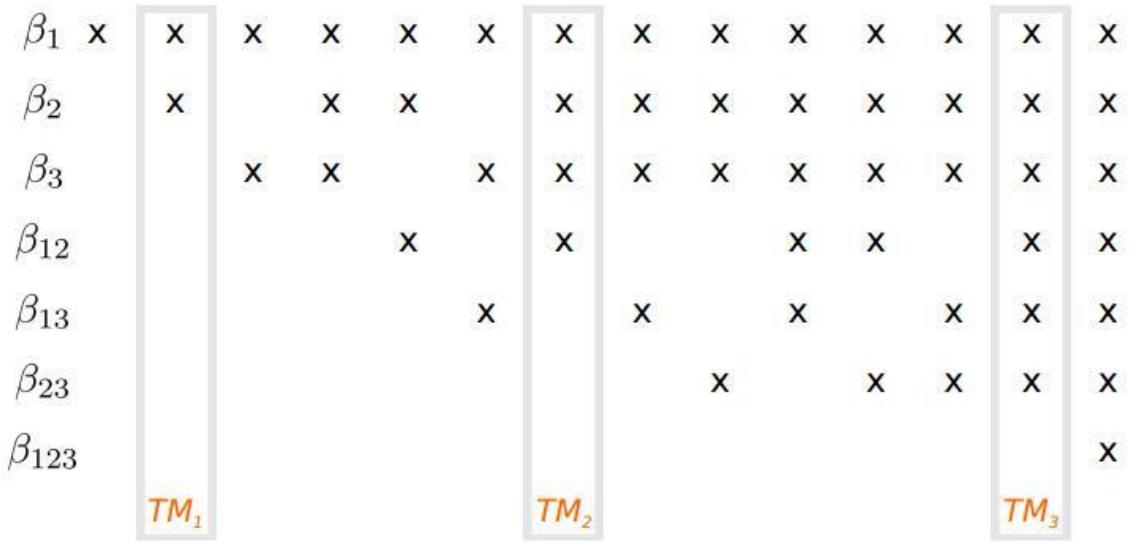
7.2.1. Models

To limit the model space within a range that is computationally feasible, we follow Devezer et. al (2019) in restraining the models to consist of a maximum of three factors. This results in 14 different combinations, which is a logical consequence of the models ranging from one to three main effects with all possible combinations including interactions. These models vary in their complexity, which increases respectively to their model number. For instance, *Model 1* is a model with only the first main effect as its predictor *($Y \sim X_1$)*, whereas *Model 14* has all three main effects and all possible interactions *($Y \sim X_1$*

$+ X_2 + X_3 + X_1{:}X_2 + X_1{:}X_3 + X_2{:}X_3 + X_1{:}X_2{:}X_3$). Model 1 and Model 14 represent the simplest and the most complex model respectively.

We choose to vary the true model between three different models. Model 2 ($Y \sim X_2$), model 7 ($Y \sim X_1 + X_2 + X_3 + X_1{:}X_2$) and model 13 ($Y \sim X_1 + X_2 + X_3 + X_1{:}X_2 + X_1{:}X_3 + X_2{:}X_3$). These models are chosen as they represent different levels of complexity and are in accordance with the true models chosen in Devezer et. al (2019) (Fig 2).

Figure 2



Taken from Devezer et. al (2019). Explaining the fourteen models. Each model is represented by a column in the matrix. Every row represents a term of the models. Every X of a column refers to what model terms are specified for that particular model. The grey areas highlight the terms of the three true models.

7.2.2. Model comparison

When it comes to model comparison, a wide variety of techniques can be employed. Each of them has their advantages and disadvantages (Emiliano, Vivanco, & de Menezes, 2014; Vrieze, 2012). As model comparison was not a main interest, we decided to use only Bayesian Information Criterion (BIC). This was done as Devezer et. al (2019) observe that iterations in which BIC is used to compare models result in more time spent at the true model than Akaike Information Criterion (AIC).

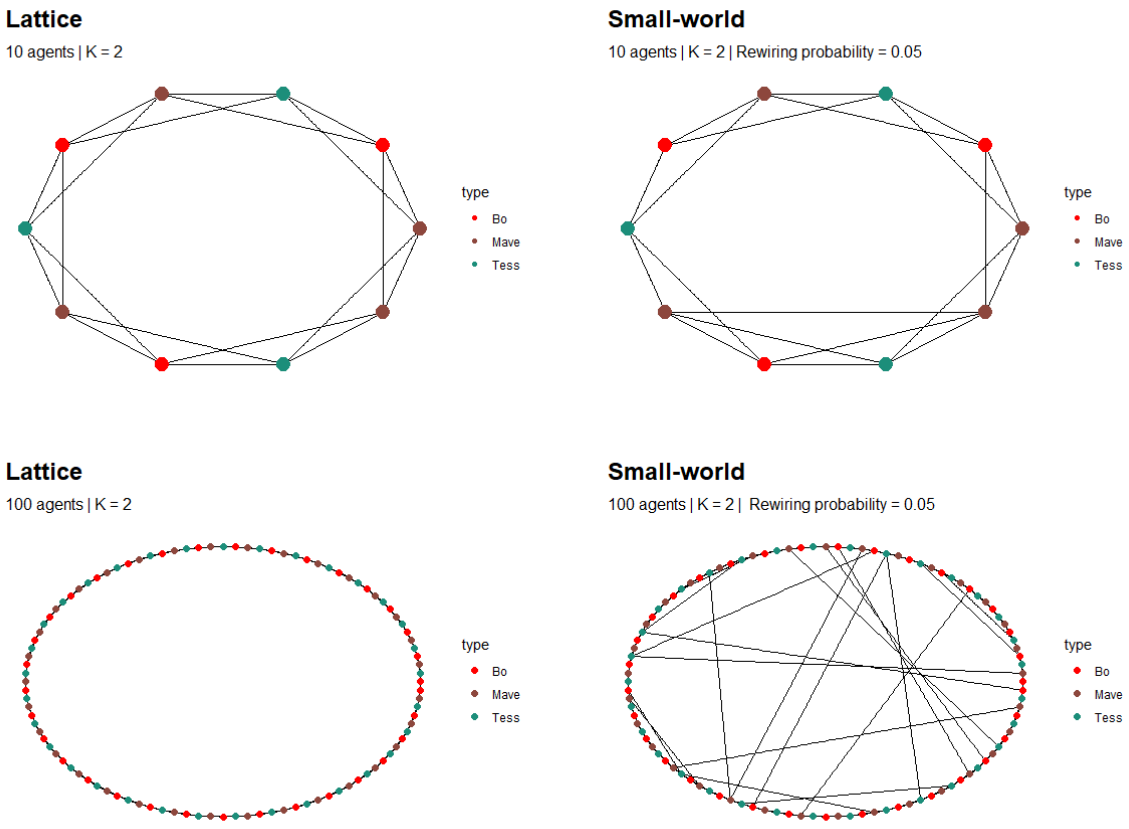
7.2.3. Networks

Networks can be either directed or undirected. In a directed network, connections between entities need not be reciprocal. We generate our network undirected, meaning that all connections are reciprocal.

The small-world structure that the agents were situated in is generated with constant settings. They all have one dimension, 100 vertices, a K-factor of 2 and a rewiring probability of 0.05. A vertice corresponds to an agent in the network. The K-factor reflects how many connections each agent has. A K-factor of 1 means that each agent has a connection to two other agents, one on each side. A K-factor of 2 means that each agent has connections to its neighbours and also its neighbours' neighbours. If the network is not manipulated further, the generated structure is called a lattice network (Fig. 3).

The small-world network diverges from the lattice by the addition of randomness. The randomness results from giving agents a rewiring probability. The rewiring probability is the probability that an agent will rewire a connection from a neighbour to another random agent in the network (Watts & Strogatz, 1998).

Figure 3



Lattice and small-world networks of different sizes. Each agent is represented as a node in the plot, and each connection is represented as a line between two agents. In the first row of the plot, a lattice (left) and a small-world network (right) is displayed, both consisting of 10 agents. For the lattice, each of the nodes have a link to their neighbour and their neighbours' neighbour. What differentiates the small-world structure from the lattice can be identified by examining the leftmost brown node. In a lattice, this node would also be connected to the red node above it. This connection has instead been rewired

across the network to the brown node on the right side of the network. In the second row of the plot, the same networks can be seen with 100 agents - the number of agents implemented in our framework.

Because of the fact that information in our network only spreads to immediate neighbours and does not cascade through the system, there should be no difference between running our simulation with a small-world structure as compared to a lattice. We chose to use the small-world network because it is reasonable to assume that scientists will differ in how well-connected they are. We run parts of our simulation with both a small-world and a lattice to make sure that they generate the same results. When results exist for both network types we will refer to comparable results of the lattice in the appendix.

7.3. Manipulations (Mikkel, Victor)

7.3.1. Sample size

We varied the sample size of the data between 100 and 20. In Devezer et. al (2019) the sample size is always set to 100 and we run simulations with this level in order to compare our results with theirs. A sample size of 20 was run as a manipulation because one of our interests is in the replication rate of psychology. Several recent papers report the median sample size of psychology and related fields to be around 20 (Marszalek, Barber, Kohlhart, & Holmes, 2011; Szucs & Ioannidis, 2017).

7.3.2. Sigma

We vary between three levels of signal-to-noise ratio; 4:1, 1:1, 1:4. These are reflected by sigmas of 0.2, 0.5 and 0.8. We inherit these levels from Devezer et. al (2019) who comment that the inclusion of low signal-to-noise ratios is reflective of psychology. Psychological research typically does not carry a strong signal, which results in low effect sizes (Szucs & Ioannidis, 2017).

7.3.3. Replications

Through our new sixth step of the process, we claim to capture replication more realistically than Devezer et. al (2019). For something to be logged as a replication study in our design two things must happen.

First, the proposed model of the agent on turn receives a lower information criterion score than her local model. This prompts her to adapt the proposed model as her new local model.

Second, an agent connected to the agent on turn shares the local model that the agent on turn had at the beginning of her turn. The agent will now do a replication study of the study just conducted by the agent on turn. Two things can now happen: This agent changes her local model because the proposed model (from the agent on turn) obtains a lower information criterion score. This represents a successful

replication. Alternatively, the agent does not change her local model because the proposed model (from the agent on turn) obtains a higher information criterion score. This represents an unsuccessful replication.

7.4. Probabilistic and random events (Victor, Mikkel)

Several steps in our simulation include probabilistic or random events. For all events which are probabilistic or random we make sure to average out the effect of randomness. First, each simulation of 11.000 studies is run 50 times. For each of these replicas we generate a new small-world network, new agents and a new sampling order. The small-world network will on average have 20 re-wirings, but some will have more and some will have less given that re-wirings are probabilistic. Because we sample new agents for each replica the agents in our network are shuffled around to different positions. Thus we do not expect any agents (and more importantly - agent types) to be consistently more well-connected than other agents. Some agents will be more connected than others in any given replica, but this should average out over repeated replicas. Lastly, the re-sampling of agents also serves to make sure that potential order effects of agents on turn is averaged out.

In addition to averaging out the effects of probabilistic and random events, we sought to make our results reproducible and comparable across different conditions. To do so, seeds were set for each of the 50 replicas.

7.4. Outcome measures (Mikkel, Victor)

To answer our hypotheses adequately, we track five different outcome measures. All measures are calculated on a replica basis and summary statistics report the median and IQR between replicas. Median and IQR were chosen to facilitate easy comparison to Devezer et. al (2019) and because we are interested in trends more than outlier cases. The measures are:

1. *Agents at the true model:* For each time-step we log the proportion of agents for which the true model is their local model. We average this proportion over the total number of time-steps per replica.

2. *50% of agents at the true model:* For each time-step we log whether more than 50% of our agent's local models were the true model. We divide the mean number of time-steps for which this is the case with the total number of time-steps per replica.

3. *Stickiness of the true model:* Stickiness of the true model is calculated as the probability that a true local model will stay true after a comparison with another proposed model.

4. *First passage time to the true model:* First passage to the true model is calculated as the first time-step where the true model is proposed and selected as the local model.

5. *Replication rate:* Replication rate represents the overall replication rate. This is calculated as the number of successful replications divided by the total number of conducted replication studies.

6. *Replication rate of the true model:* This measure is calculated as the number of successful replications of the true model divided by the total number of conducted replication studies of the true model. This will be a subset of the total number of replication studies.

8. Results (Mikkel, Victor)

We now turn to the results of our ABM. In this section we will start by comparing the three main findings of Devezer et. al (2019) to ours, and then we will turn to our two further hypotheses. This results in a total of five sections. The discussion will refer to results from simulations done on a small-world structure and a sample size of 100 when nothing else is specified. These settings provide the baseline which we will at times compare to results from our ABM on a lattice network and on a small-world network with sample size 20.

Figure 4

	Small-world | Sample size: 100							
	Median				IQR			
	All	Bo	Mave	Tess	All	Bo	Mave	Tess
Agents at the true model	29.76	4.11	62.00	13.75	27.86	3.34	66.53	41.89
50% of agents at the true model	0.00	0.00	84.29	0.00	0.00	0.00	90.18	27.01
Stickiness of the true model	96.82	96.45	96.25	98.79	27.74	3.16	22.55	47.97
First passage time to the true model	16.00	77.00	10.00	16.00	28.00	219.00	19.00	35.75
Replication rate	39.23	17.78	31.42	32.48	15.15	21.55	9.51	28.53
Replication rate of the true model	62.16	58.71	63.01	54.60	27.11	45.73	22.51	54.88

Note:
Summary statistics for all populations. First passage time to true model in number of time-steps. All else in percentage points. Averaged over iterations, sigmas and true models

8.1. The benefits of diversity (Victor, Mikkel)

Recall our brief discussion of the benefits of diversity from the introduction. In that section we singled out two key features which favour diversity: averaging and diminishing returns to type. Devezer et. al (2019) mainly focus on the benefits of averaging in their discussion. They conclude that all homogenous populations perform poorly on at least one measure of scientific success while the epistemically diverse population is never the worst on any outcome measure (Devezer et. al, 2019). The same trend is qualitatively replicated in our ABM. *Bo*-dominant populations spend the least time at the true model, have the slowest first passage time to the true model, and also the lowest replication rate. *Tess*-dominant populations have the lowest replication rate of the true model, and in addition they have most variability on replication rate of the true model, replication rate and stickiness of the true model. *Mave*-dominant populations do well on many measures but do show highest variability of agents at the true model. In contrast the epistemically diverse population never does worst on any outcome measure. It almost never attains consensus, what we have operationalized as at least 50% of agents having the true model as their local model. However, the widespread lack of consensus is shared by all other populations except for the *Mave*-dominant one. For most outcome measures the epistemically diverse population has the second highest median score and the second lowest interquartile range (Fig. 4).

This result is unsurprising as the epistemically diverse population is comprised of a mixture of the other agent types and should thus be "dragged down" by those who do worst on any outcome measure, while being "pulled up" by those agent types who perform well.

However, in addition to the benefit of averaging we ask the additional question of whether there is an over-and-above benefit to diversity, resulting from diminishing returns to type or by other means. Here the picture is less clear. The epistemically diverse population does better than the mean of the homogenous populations on first passage time to true model, and marginally better than the mean of the homogenous populations on agents at the true model. However, the epistemically diverse population never does best on any of the desirable properties of science. The epistemically diverse population has the highest replication rate, but as we have argued replication rate is not an intrinsically desirable property of science.

We must thus conclude that our ABM shows an averaging effect of diversity, but that there is not a clear net benefit to diversity in our system. The same pattern is found in a small-world with a sample size of 20, and in a lattice (App. A & B). We believe that our inability to generate a clear case of diminishing returns to type results from the constraints we placed on information sharing in our system. Think about it this way: If a well-performing agent-type in a certain condition finds the true model then the neighbours of this agent will be likely to obtain this model as their local model as well. This is what communication does in our network. However, these agents are now very unlikely to change their local

model when they are on turn because they already have the true model as their local model. We can see this from the very high stickiness of the true model across populations (Fig. 4). The result is that their neighbours will not be alerted that they have found a good model. They will not benefit from the finding of the original agent. There will be no rippling effect through the system and thus the marginal value that any agent can confer to the system is limited. We believe that a design change of the ABM which allows for information to cascade through the system would be likely to produce effects of diminishing returns to type and thus highlight the benefits of diversity more clearly than what we see here.

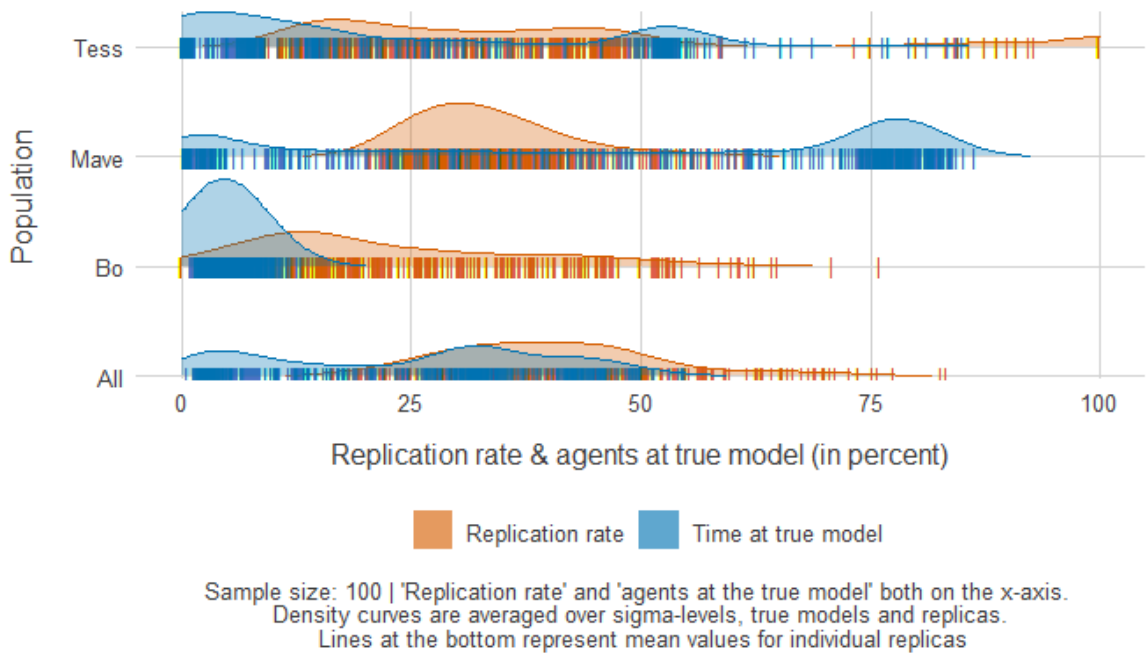
8.2 Innovation drives the speed of progress (Mikkel, Victor)

Innovation in our system is embodied by the agent type *Mave* as explained earlier. In Devezer et. al (2019) Mave populations consistently had the lowest first passage time to the true model. This finding is replicated in our system, where *Mave*-dominant populations reach the true model faster than any other population (Fig. 4). The finding is robust across conditions, and thus also replicates in the lattice and in the small-world in which agents conduct studies with a sample size of 20 (appendix A & B). We will explore and discuss the properties of the agent type *Mave* further in the section concerning consensus.

8.3 Replication rate & desirable scientific properties (Mikkel, Victor)

In Devezer et. al (2019) they argue that there is no *causal* relation between replication rate and other desirable scientific properties. They also show that there is in fact no correlation between time spent at the true model and replication rate. This is a surprising finding given our earlier argument that replications are usually taken to indicate truth (Goodman et al., 2016). Contrary to Devezer et. al (2019) we do find small correlations between replication rate and desirable properties of science, but these associations provide a murky picture that is hard to interpret (Fig. 4 & Fig. 5).

Figure 5



We find a positive correlation between replication rate and agents at the true model ($r_{SR} = 0.18$) and between replication rate and stickiness of the true model ($r_{SR} = 0.15$). This might seem encouraging, but we also observe a negative correlation between replication rate and 50% of agents at the true model ($r_{SR} = -0.14$) and a negative correlation between replication rate and first passage time to true model ($r_{SR} = -0.25$). Because these correlations are relatively weak and inconsistent in their direction we choose not to interpret them individually. Instead we note the lack of a consistently positive association between replication rate and desirable scientific properties. To us this qualitatively replicates the finding from Devezer et. al (2019).

The dissociation between desirable scientific properties of science and replication rate becomes even more salient when we run our ABM with a sample size of 20. In this condition the populations spend much less time at the true model and they also show much less stickiness of the true model. Surprisingly, the replication rate stays approximately the same as in the conditions with sample size 100 (compare Fig. 5 & App. C).

8.4 Low replication rate (Victor, Mikkel)

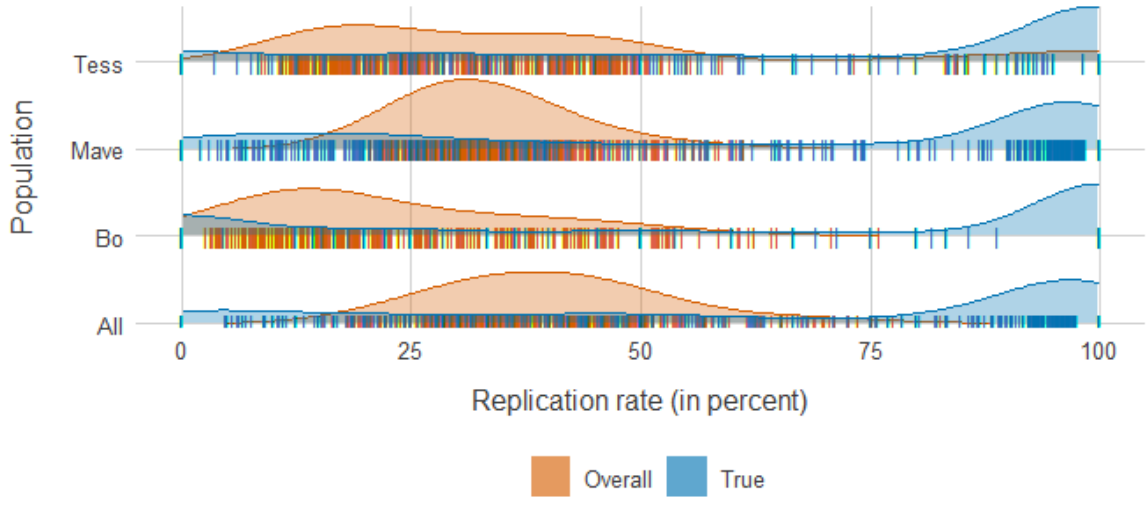
As discussed earlier low replication rates are the norm in the social sciences, most notably in psychology (A. Aarts et al., 2015). This is a macrostructure that we expect our model to generate. The replication

rates in our ABM range from 17.78 to 39.23% but is generally in the thirties. Only *Bo*-dominant populations perform somewhat worse (Fig. 4).

These replication rates are much lower than the ones observed in Devezer et. al (2019) who report replication rates between 89.9 and 100%. Our replication rates resemble those reported in the large-scale attempt to replicate findings in psychology which was discussed earlier (A. Aarts et al., 2015). As noted however, these quantitative comparisons should not be given too much weight given the crude nature of our system. However, we are confident that our system does provide a more realistic picture of replication rates than that of Devezer et. al (2019). A realistic replication rate lends further credibility to the claim made in the previous section that there is no consistently positive correlation between replication rate and desirable properties of science.

Interestingly, we find a substantial difference between the overall replication rate and the replication rate of the true model. Replication rates for the true model figures at around 50 to 60% in our ABM, while overall replication rates generally figure around 30% (Fig. 4 & Fig. 6). This discrepancy was not found in Devezer et. al (2019), perhaps because of a ceiling effect (I.e., several populations had a replication rate of the true model of 100%).

Figure 6



Sample size: 100 | Density curves are averaged over sigma-levels, true models and replicas.
Lines at the bottom represent mean values for individual replicas.
Outliers (I.e., replication rates of 0% or 100%) occur because some replicas contain very few replication studies. Some replicas have no replication studies of the true model. The overall picture however remains clear.

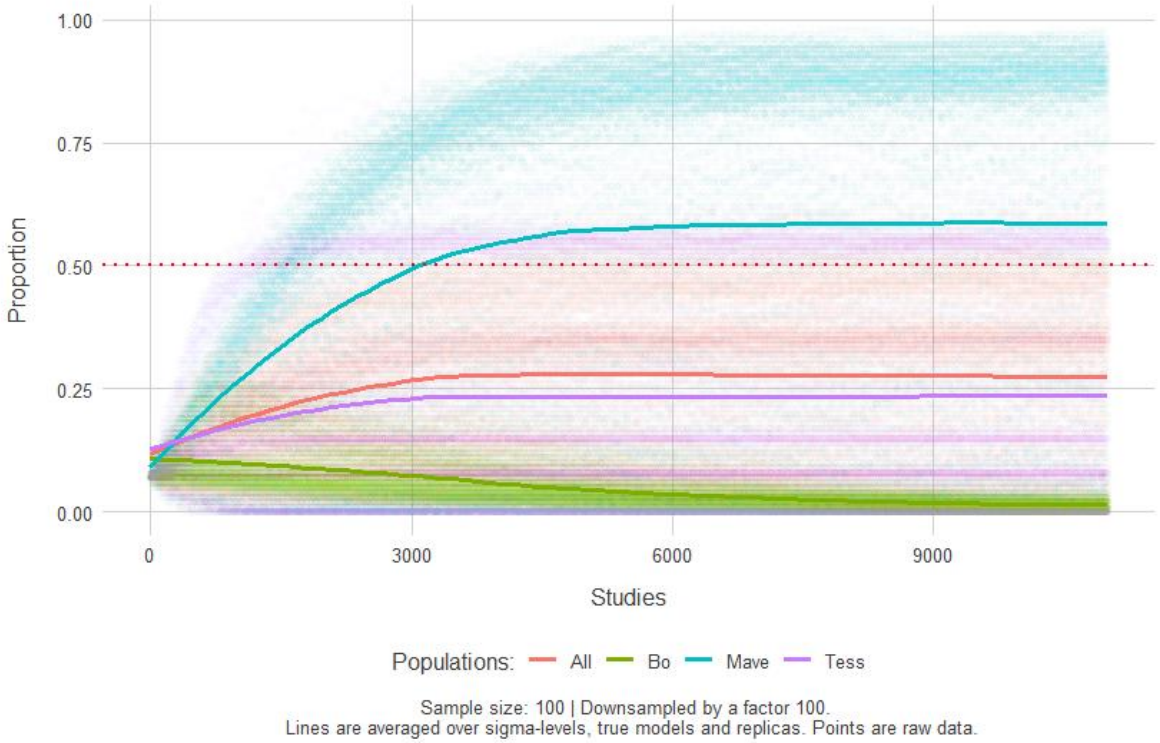
The fact that we observe a sizeable difference between the two measures indicates that replicability could be a useful tool for the scientific community. We retain that the overall replication rate of a field does not appear to be clearly associated with the proportion of scientists in that field who adhere to a

true model. On a case-by-case basis however, replication might still serve as our best indicator of truth. If possible candidates for a true model receive multiple replication attempts, it should be possible to distinguish fairly consistently between true and false findings.

8.5. Reaching consensus (Mikkel, Victor)

We now zoom in on scientific consensus. The most salient result here is that *Mave*-dominant populations reliably reach consensus on the true model (Fig. 4 & Fig. 7). Mave does not only outperform her counterparts; she does surprisingly well. During some conditions, *Mave*-dominant populations reach up to 90% consensus on the true model (see the raw data in Fig. 7). Despite all the other populations having a median of zero on our consensus measure, they vary considerably (Fig. 4). *Tess*-dominant populations consistently do well when the true model is model 2, which is the simplest of our true models (notice the IQR in Fig. 4 & see Fig. 7). Interestingly, *Tess*-dominant populations do worse on average than the epistemically diverse population. The epistemically diverse population often stabilize at around 20-50% of agents adhering to the true model (Fig. 7). *Bo*-dominant populations spend very little time at the true model and we have recorded no replicas in which she reaches consensus on the true model. *Bo*-dominant populations prefer the most complex model no matter what the true model is (App. D).

Figure 7



One of the most surprising findings is the overall high performance-level of *Mave*-dominant populations. Multiple explanations can be given to why this is the case. First, it is important to realize that Mave's research strategy can be considered inherently diverse. *Mave*-dominant populations are classified as homogenous in our system because they consist largely of the same type. However, because she chooses randomly she can be interpreted as embodying a mixture of several research strategies.

As Devezer et. al (2019) note, the constrained model space which we have adapted from them might wrongly indicate that random and unbiased sampling is always the best research strategy. In the real world we do not have the luxury of knowing that one of 14 models is guaranteed to be true, so the potential model space is effectively infinite. In that case one needs to somehow constrain one's search for truth based on reasonable heuristics or contend with being highly inefficient. Our results do capture some of this inefficiency. *Mave*-dominant populations show much more variance than the epistemically diverse population in the 'agents at the true model' measure, as indicated by higher IQR scores (Fig. 4).

It might seem surprising at first that we almost never see the epistemically diverse population reach consensus. However, this is due to the effect of averaging discussed earlier. While *Mave*- and *Tess*-dominant populations do reach consensus under certain conditions, it is typically the case that the conditions under which one agent type flourishes sees the other agent types perform poorly. In addition, *Bo*-dominant populations always perform poorly. Because the epistemically diverse population averages both high-performers and low-performers it consistently produces mediocre outcomes.

In sum, the only population which defies our expectations and consistently reaches a consensus around the true model is *Mave.* However, it must be noted that when we change the sample size from 100 to 20 we do not see any populations consistently reaching consensus (App. A & App. E). The scientific fields in which a lack of consensus is dominant are also the social sciences in which sample sizes are typically closer to 20 than 100 (Collins, 1994; Marszalek et al., 2011; Szucs & Ioannidis, 2017).

9. Limitations (Victor, Mikkel)

9.1. Microspecifications (Victor, Mikkel)

We now consider a couple of general unsatisfactory features of our ABM which could be improved. The first issue concerns memory in the system. Devezer et. al (2019) note the lack of memory in their system as a major flaw. Our ABM did not remedy this, and we believe it to be one of the biggest issues in our system. The unfortunate consequence of not including memory in the system is that every time a proposed model does better than the local model of an agent, the agent will adopt the proposed model.

We believe that scientists in the real world cling more stubbornly to their ideas and in some cases for good reasons. It seems unrealistic to assume that a scientist who has consistently found that model "x" performs better than model "y" would suddenly change her mind when seeing model "y" outperform model "x" once. The difficulty that Devezer et. al (2019) note lies in deciding on a satisfactory decision rule for when an agent in the system should change her mind. In this regard, it is necessary for the scientist employing the ABM to interact with behavioural researchers to obtain a realistic microspecification (Epstein, 1999).

Another issue concerning the behavioural rules of agents is how they sample new models. In other words, the extent to which agent types in our system correspond to real-world research strategies is questionable. The usefulness and realism of the microspecification for the agent type *Bo* appears most problematic. *Bo* is designed to represent a boundary testing approach to science. This strategy is not uncommon in the scientific community (Gonzalez-Mulé & Aguinis, 2018; Whetten, 1989). However, it seems unrealistic to always and unconditionally propose more complicated models. At least we can say that *Bo*-dominant populations consistently perform poorly in our ABM (Fig. 4, App. D). In general, the specifications which define agent types should be informed by more empirical research into how scientists choose to test what they do.

9.2. Information cascades (Mikkel, Victor)

One of the aims of representing agents in a network was to constrain how information is shared among members of a community. We believe that it was important to limit the information in the network. However, as was briefly discussed in the section on the benefits of diversity our design resulted in information being too constrained. We believe that it is realistic that no scientist reads everything or communicates with every other scientist. However, it is the case that when new and convincing studies are done they ripple throughout scientific communities (Young, Ioannidis, & Al-Ubaydli, 2008). Allowing information to cascade through the system would also make research on the efficiency of different network structures in the scientific community possible. As has been noted our very limited communication has the effect that the only relevant properties of the network is its size and the number of connections. The effects of how scientists cluster, and how fast information can travel far in the network are not possible to investigate given the current model design.

10. Conclusion

Our ABM replicates the findings from Devezer et. al (2019) concerning the averaging effects of diversity. We also replicate the finding that innovation is a desirable property of scientific research. The ecological validity of this finding was questioned on the grounds that our model space is very limited and might indicate innovative approaches to be more fruitful than they are in the real world. Finally, we replicate the finding that replication rate is not consistently associated with other desirable properties of a scientific field. This surprising finding is strengthened by the fact that our replication rate quantitatively differs from theirs but still exhibits no predictive value against desirable scientific properties.

In contrast to the original findings, we find that true findings have a replication rate which is almost twice as high as the overall replication rate. This indicates that replications can have diagnostic value on a case-by-case basis. Moreover we find that scientific consensus around truth is rare, and never obtained when we adjust the sample size to match median sample sizes in the psychological literature.

Our model is still crude, but our additions to the model used in Devezer et. al (2019) expand and improve on the framework. This is evident from a more representative replication rate and more ecologically valid assumptions. We hope that others will develop the framework further, perhaps so as to investigate the effects of different network structures or by incorporating memory into the system.

11. Acknowledgements

We would like to thank Devezer et. al (2019) for allowing us to expand upon their framework and code.

Their code is freely available on Github: https://github.com/gnardin/CRUST/tree/master/src/abm

Our code is freely available on Github: https://github.com/teamsoccult/CRUST-1

12. Appendix

Appendix A

	Small-world | Sample size: 20							
	Median				IQR			
	All	Bo	Mave	Tess	All	Bo	Mave	Tess
Agents at the true model	12.00	1.26	19.73	7.61	23.52	0.78	36.56	20.50
50% of agents at the true model	0.00	0.00	0.00	0.00	0.00	0.00	17.55	0.00
Stickiness of the true model	74.21	83.54	78.78	72.16	39.96	11.27	42.14	50.70
First passage time to the true model	19.00	67.50	16.00	18.00	31.00	319.00	20.00	41.75
Replication rate	36.32	33.33	35.65	31.02	13.51	19.17	7.92	17.58
Replication rate of the true model	56.70	66.67	66.81	40.00	21.32	50.00	11.62	34.35

Note:
Summary statistics for all populations. First passage time to true model in number of time-steps. All else in percentage points. Averaged over iterations, sigmas and true models.

Appendix B

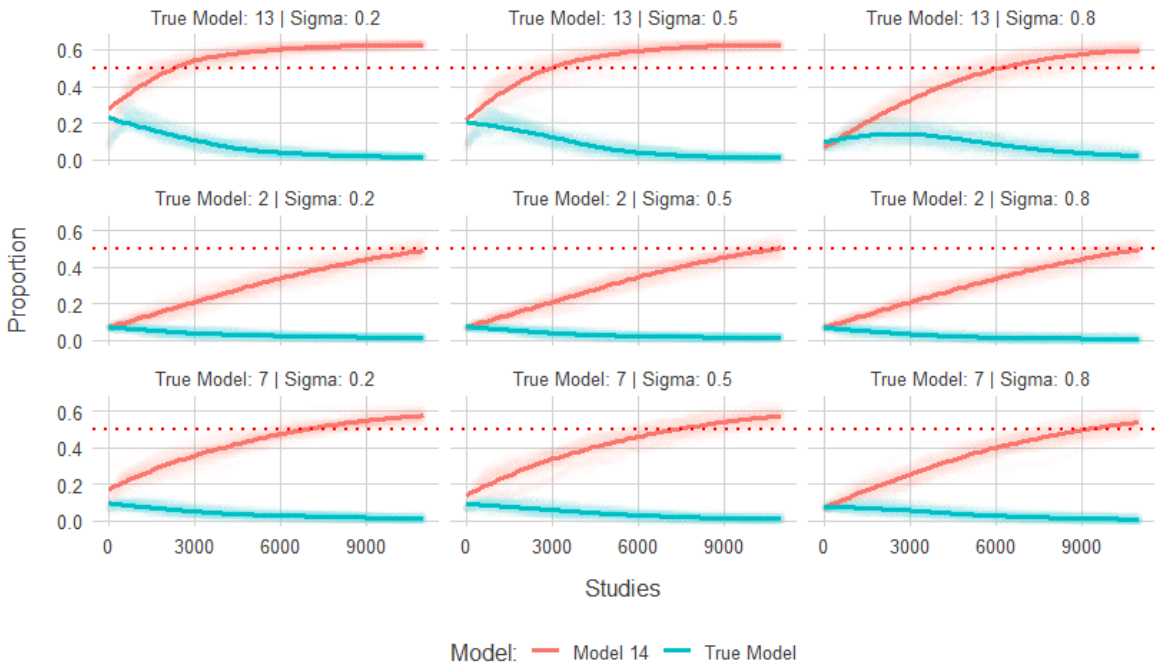
	Lattice | Sample size: 100							
	Median				IQR			
	All	Bo	Mave	Tess	All	Bo	Mave	Tess
Agents at the true model	29.10	4.23	59.69	12.95	29.13	3.49	66.78	36.90
50% of agents at the true model	0.00	0.00	84.22	0.00	0.00	0.00	90.14	4.38
Stickiness of the true model	95.43	96.30	95.87	98.82	26.50	3.25	25.32	49.72
First passage time to the true model	14.00	80.50	12.00	20.00	26.00	338.25	17.50	45.00
Replication rate	40.37	18.68	32.42	30.95	15.50	21.86	10.77	28.49
Replication rate of the true model	62.92	55.56	63.12	55.93	24.85	38.75	19.05	74.69

Note:
Summary statistics for all populations. First passage time to true model in number of time-steps. All else in percentage points. Averaged over iterations, sigmas and true models.

Appendix C

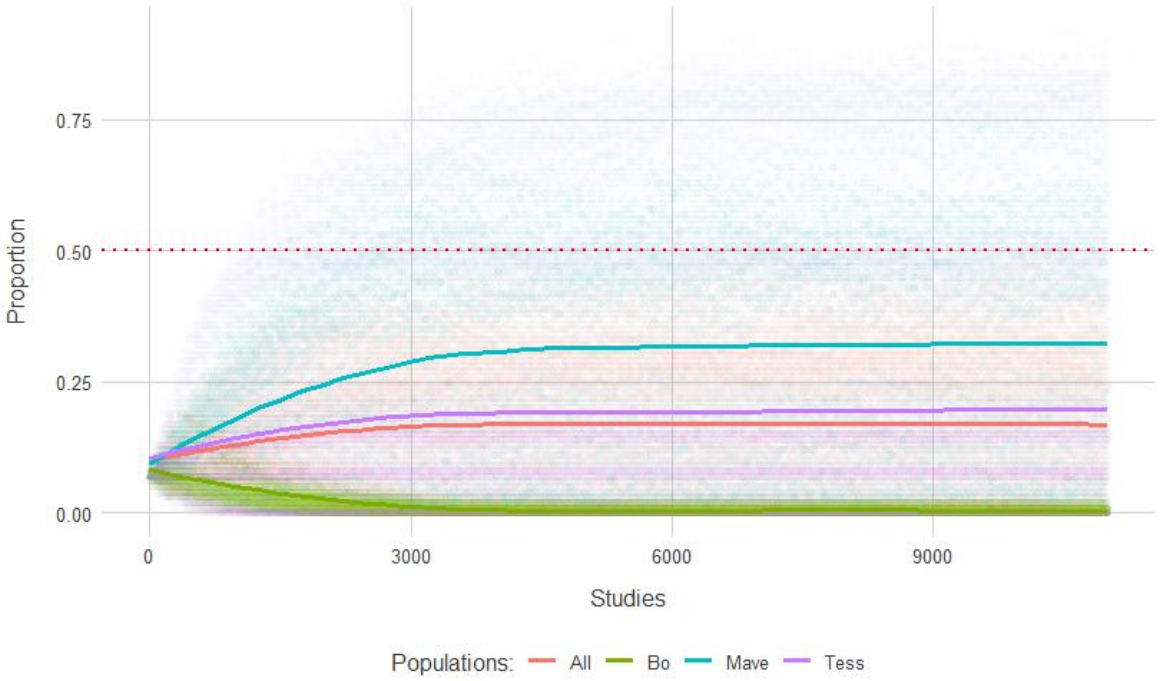
Sample size: 20 | 'Replication rate' and 'agents at the true model' both on the x-axis.
Density curves are averaged over sigma-levels, true models and replicas.
Lines at the bottom represent mean values for individual replicas

Appendix D

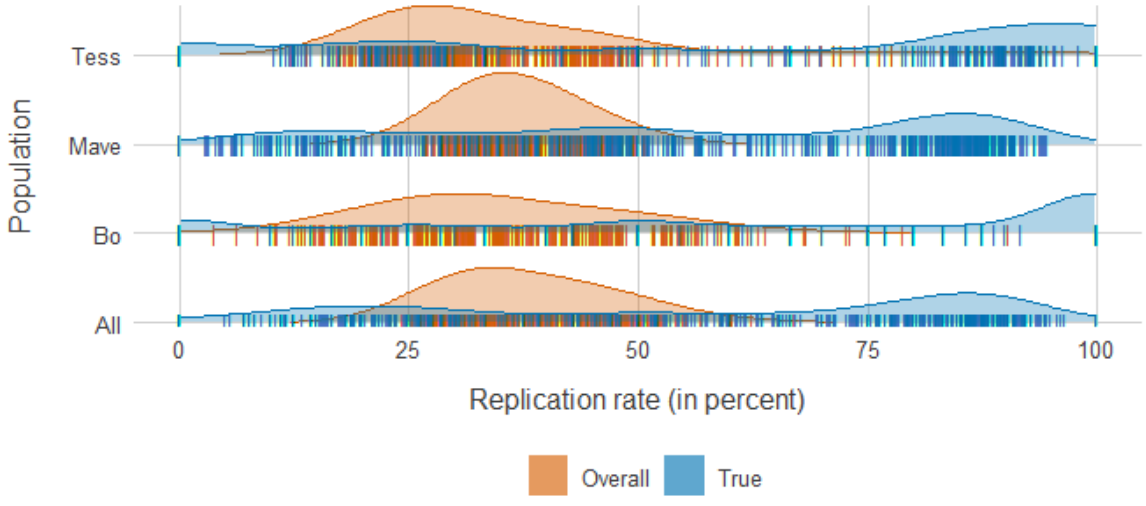


Sample size: 100 | Population: Bo | Downsampled by a factor 100.
Lines are averaged over replicas. Points are raw data.

Appendix E



Appendix F



13. Literature

A. Aarts, A., E. Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., … Penuliar, M. (2015). Estimating the Reproducibility of Psychological Science. *Science*, *349*. https://doi.org/10.1126/science.aac4716

Collins, R. (1994). Why the social sciences won't become high-consensus, rapid-discovery science. *Sociological Forum*, *9*(2), 155–177. https://doi.org/10.1007/BF01476360

Crooks, A. T., & Heppenstall, A. J. (2012). Introduction to Agent-Based Modelling. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-Based Models of Geographical Systems* (pp. 85–105). https://doi.org/10.1007/978-90-481-8927-4_5

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), e0216125. https://doi.org/10.1371/journal.pone.0216125

Ebadi, A., & Schiffauerova, A. (2015). On the Relation between the Small World Structure and Scientific Activities. *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0121129

Emiliano, P. C., Vivanco, M. J. F., & de Menezes, F. S. (2014). Information criteria: How do they behave in different models? *Computational Statistics & Data Analysis*, *69*, 141–153. https://doi.org/10.1016/j.csda.2013.07.032

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60. https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F

Gonzalez-Mulé, E., & Aguinis, H. (2018). Advancing Theory by Assessing Boundary Conditions With Metaregression: A Critical Review and Best-Practice Recommendations. *Journal of Management*, *44*(6), 2246–2273. https://doi.org/10.1177/0149206317710723

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

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

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Page, S. E. (2010). *Diversity and Complexity*. Princeton University Press.

Smaldino, P. E. (2017). Models Are Stupid, and We Need More of Them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (1st ed., pp. 311–331). https://doi.org/10.4324/9781315173726-14

Smaldino Paul E., & McElreath Richard. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440. https://doi.org/10.1038/30918

Whetten, D. A. (1989). What Constitutes a Theoretical Contribution? *Academy of Management Review*, *14*(4), 490–495. https://doi.org/10.5465/amr.1989.4308371

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. *PLOS Medicine*, *5*(10), e201. https://doi.org/10.1371/journal.pmed.0050201