

Not Null Enough: Pseudo-Null Hypotheses in Community Ecology and Comparative Psychology

William C. Bausman

Marta Halina

Forthcoming in *Biology & Philosophy*, accepted 20/8/18

Abstract:

We evaluate a common reasoning strategy used in community ecology and comparative psychology for selecting between competing hypotheses. This strategy labels one hypothesis as a “null” on the grounds of its simplicity and epistemically privileges it as accepted until rejected. We argue that this strategy is unjustified. The asymmetrical treatment of statistical null hypotheses is justified through the experimental and mathematical contexts in which they are used, but these contexts are missing in the case of the “pseudo-null hypotheses” found in our case studies. Moreover, statistical nulls are often not epistemically privileged in practice over their alternatives because failing to reject the null is usually a negative result about the alternative, experimental hypothesis. Scientists should eschew the appeal to pseudo-nulls. It is a rhetorical strategy that glosses over a commitment to valuing simplicity over other epistemic virtues in the name of good scientific and statistical methodology.

Acknowledgements:

Versions of this paper were presented at POBAM 2014 and SPSP 2015. We thank the audience for their questions and discussion. We would also like to thank Adrian Currie, Shay Logan, Helen Longino, Elliott Sober, Kent Staley, Jos Uffink, and C. Kenneth Waters for extensive comments and discussion. William Bausman's writing of this article was supported in part by a grant from the John Templeton Foundation: #50191; *From Biological Practice to Scientific Metaphysics*.

1. Introduction

A central goal in science is identifying and justifying the best explanation for a given phenomenon. There are many strategies for doing this, such as evaluating the empirical adequacy of the available hypotheses and assessing their relative epistemic virtues. In this paper, we examine a strategy for selecting between competing scientific hypotheses. A central feature of this strategy is using what we call a “pseudo-null.” This is a hypothesis that practitioners in the field label a null hypothesis, but which lacks the features of a true statistical null. A true statistical null is used within the context of a well-designed experiment and it is within this context that inferences concerning whether to retain or reject the hypotheses under consideration are supported. A core condition of the inference strategy in statistical hypothesis testing is that one cannot accept the alternative hypothesis until one has rejected the null hypothesis. In the cases we consider here, this inference strategy is also employed, but it is done so outside the context of statistical hypothesis testing. Instead, proponents of this strategy label one hypothesis the “null” on the grounds of its simplicity and argue that any alternative hypothesis cannot be accepted until researchers reject this null. Insofar as the null cannot be rejected, the proponents argue, it should be accepted as the best explanation of the phenomenon under investigation. In this paper, we argue that this pseudo-null strategy is unjustified and should be abandoned.

We begin in section 2 by showing how the appeal to pseudo-nulls is used in community ecology and comparative psychology. In community ecology this strategy is used to defend the neutral theory of ecology, while in comparative psychology it is used to defend the behavior-reading account of social behavior in nonhuman animals. Proponents of both hypotheses characterize these hypotheses as “nulls” on the grounds that they are simpler than the available alternatives. In section 3, we show that this form of argument is superficially similar to the

statistical method of Neyman-Pearson testing, thus giving it rhetorical force, but go on to argue that it is disanalogous in precisely those respects required for justification. Whereas the null hypotheses used in Neyman-Pearson testing concern the effects of controlled or randomized extraneous variables, the pseudo-null hypotheses in our case studies are used outside this context. We also extend our argument to counter the objection that model selection theory can justify the pseudo-null strategy and consider several additional possible justifications. We conclude in section 4 that pseudo-nulls should be treated on a par with the available alternative hypotheses and the rhetoric of testing null hypotheses should be dropped. In the end, we hope our analysis will stimulate critical discussions on the role that nulls play more broadly in the sciences.

Before we begin, it is important to clarify the target of our critique. First, we do not aim to critique the specific hypotheses advanced by the pseudo-null strategy, but rather the strategy itself. There may be better arguments or alternative lines of evidence favoring the behavior-reading hypothesis and neutral theory; we do not deny that this is the case. Second, we do not offer here arguments against all possible justifications for the pseudo-null strategy. Instead, we argue that its justification does not come from an analogy to statistical inference. We focus on statistical inference because we think this is where the pseudo-null strategy gets its rhetorical force. Users of this strategy do not give grounds for its use, but rather employ the language of “null hypothesis testing” and the associated inferential strategy as if its justification were clear. The strategy looks unproblematic and without need for explicit justification, we believe, because it superficially resembles statistical inference. Thus, our aim here is to show that the pseudo-null strategy does not in fact share justificatory grounds with statistical inference. The onus falls on users of the pseudo-null strategy to explain its justification. Also, the language of “null

hypothesis testing” should be avoided, as it is misleading. It leads one to think this is a straightforward form of statistical inference when it is not.

2. The Pseudo-Null Strategy

The pseudo-null strategy advances a particular explanatory hypothesis by appealing to it as a “null,” where its status as a null is justified on the grounds of simplicity. Rather than appealing to simplicity as an epistemic virtue, however, this approach epistemically privileges hypotheses in the following way: it holds that (1) one must reject the null before one is warranted in accepting any alternative explanations and (2) if the null cannot be rejected, it should be accepted as the best explanation for the phenomenon under study. In what follows, we illustrate how this approach is used in community ecology and comparative psychology.

2.1. Explaining Relative Species Abundance Distributions in Community Ecology

A long-standing problem within ecology is to explain patterns of diversity and abundance of species within communities. A particular pattern of interest is the number of species found within a single trophic level and their population sizes—that is, distributions of relative species abundance. The traditional and still dominant approach to explaining this phenomenon holds that abundances result from competition between species for resources and tradeoffs between how different species utilize those resources (Chase and Leibold 2003). For example, the competitive exclusion principle holds that, if species coexist, then there are differences between how the species utilize resources. Mechanisms that depend upon species differences, including the mechanism of interspecific competition, are called *selection* mechanisms by analogy to natural selection (Vellend 2016).

Stephen Hubbell has challenged the selection hypothesis by advancing the neutral theory of ecology—a theory that is now both prominent and controversial in the discipline (Hubbell 2001, 2006). According to neutral theory, communities are structured entirely by three mechanisms: ecological drift, random immigration, and random speciation. The theory is *neutral* because it assumes all individuals, regardless of species, have identical chances of giving birth, dying, immigrating, and being a member of a new species. Neutral theory predicts that a community’s observed relative species abundance distributions will fit a particular statistical distribution. The free parameters tuning the distribution are interpreted in terms of community size, metacommunity size, immigration rate, and speciation rate.

The neutral theory is explicitly constructed in order to exclude the influence of selection—for this reason it is said to be simpler than the selection hypothesis. Hubbell writes that the value of constructing the neutral theory of ecology is that, “we obtain a quantitative null hypothesis against which to test when, to what extent, and for which species demographic differences among species are necessary to explain observed community patterns” (Hubbell 2006, 1387). The theory is not neutral in the sense of “nothing going on,” but rather in excluding species differences as causally responsible for the species abundance distributions observed in a particular community. Because selection depends on species differences, assuming neutrality is taken to justify using neutral theory to supply the appropriate “null hypothesis” for testing selection hypotheses.

2.2. Explaining Chimpanzee Social Behavior in Comparative Psychology

Great apes such as chimpanzees engage in sophisticated social behavior. When stealing food from a competitor, for example, a chimpanzee will prefer to steal food that a competitor

cannot see (Hare, Call, and Tomasello 2006, Melis, Call, and Tomasello 2006). The dominant explanation for such behavior is that chimpanzees track the mental states of others: they know (consciously or not) when other agents can or cannot see particular objects in their environment. Comparative psychologists currently hold that chimpanzees reason about a variety of mental states such as goals, intentions, and perceptions (Call and Tomasello 2008). In accordance with human psychological research, this ability is referred to as “mindreading.”

Daniel Povinelli and colleagues have challenged the above consensus by arguing that there is a competing explanation for the apparent mindreading abilities of chimpanzees (Penn, Holyoak, and Povinelli 2008, Penn and Povinelli 2007, Povinelli and Vonk 2004, Penn and Povinelli 2009). Their “behavior-reading” hypothesis holds that chimpanzees employ a set of abstract behavioral rules that allow them to anticipate how other agents will behave based on prior behavior and environmental circumstances. On this view, the fact that chimpanzees prefer to steal food that a competitor cannot see is best explained by the implementation of some rule, such as, ‘if there is an unobstructed line of gaze from an agent’s eye to a piece of food, then that agent will prevent the food from being stolen’.

Proponents of behavior-reading refer to this hypothesis as a “null” on the grounds that it is simpler than the mindreading hypothesis. On the mindreading account, a chimpanzee predicts the behavior of other agents based on their mental states (e.g., they cannot see x), which is inferred from their observable behavior (such as the fact that their head is turned away from x). The behavior-reading hypothesis, in contrast, holds that chimpanzees do not perform the inferential step involving positing a mental state, but rather make their behavioral predictions based on the observable behavior alone (Penn and Povinelli 2007, Lurz 2011).

2.3. The Asymmetrical Evaluation of Hypotheses

The neutral theory in community ecology and the behavior-reading hypothesis in comparative psychology are presented as null hypotheses by their proponents. The justification given for their status as nulls is that they are simpler than the competing alternative hypotheses. Crucially, treating these hypotheses as nulls drastically changes how they are evaluated. Typically, two competing hypotheses are evaluated by comparing their empirical adequacy and theoretical virtues (such as internal consistency, coherence with other theories, generality, etc.). In contrast, the neutral theory and behavior-reading hypotheses enjoy a significant epistemic advantage. Proponents of these hypotheses hold that insofar as they are empirically adequate, they should be accepted as the best explanation of the phenomenon of interest. Unless the null can be shown to be empirically inadequate, all empirically adequate alternatives must be rejected regardless of their theoretical virtues. Anyone advocating for one of the competing alternative hypotheses bears the burden of disproving the null.

This asymmetrical treatment of hypotheses is exhibited in both of our case studies. In the case of community ecology, Hubbell maintains that, “obtaining acceptable fits from neutral models shifts the burden of proof to those who would assert that more complex theory is required to explain nature and with what level of detail and generality” (Hubbell 2006, 1387). Because the neutral theory is simpler than theories including species differences, the burden is on advocates of the more complex theory to disprove the neutral theory. Hubbell argues that the best explanation for why the neutral theory fits the data well is that “it accurately captures some . . . characteristic of the behavior of biodiversity in aggregate and at large spatial and temporal scales” (Hubbell 2006, 1388). Until the neutral theory can be rejected, the processes it posits should be used to explain the patterns it fits well. And so drift, immigration, and speciation

explain species abundance distributions when the fit is good.

Proponents of the behavior-reading hypothesis also hold that it must be falsified before one is justified in accepting the claim that nonhuman animals mindread. As Penn and Povinelli write, “in order to produce experimental evidence for an f_{ToM} [theory of mind function], one must first falsify the null hypothesis that the agents in question are simply using their normal, first-person cognitive state variables” (Penn and Povinelli 2007, 734). According to Povinelli and colleagues, until the behavior-reading hypothesis is rejected, we have “no evidence for theory of mind in animals” (Penn and Povinelli 2007, 732). Moreover, if the behavior-reading hypothesis cannot be rejected, we should accept it as the best explanation of nonhuman animal social behavior. Thus, Penn and Povinelli conclude, “the available evidence suggests that chimpanzees, corvids and all other non-human animals only form representations and reason about observable features, relations and states of affairs from their own cognitive perspective” (Penn and Povinelli 2007, 737).

The pseudo-null strategy is similar to one that Elisabeth Lloyd discusses in the context of adaptationism in evolution (Lloyd 2015). Lloyd identifies cases where proponents of one scientific hypothesis will cast a competing hypothesis as a mere statistical “null hypothesis” to undermine their competition. This casts the 'null' as a false hypothesis that should be rejected if the science is going well. The effect is to illegitimately privilege the non-“null” hypothesis. In the pseudo-null strategy however, it is the “null” hypothesis that is privileged. We think both strategies are unjustified for the same reasons and our analysis here complements Lloyd's by critically assessing the use of “null” language in other areas of science.

3. Why the Pseudo-Null Strategy Lacks Justification

The above asymmetrical treatment of hypotheses has important consequences on research. Rather than weighing the theoretical virtues of two empirically adequate hypotheses, the pseudo-null strategy prompts one to accept the null hypothesis when you cannot reject it. In this section, we consider the justification for this strategy. We show that the approach resembles Neyman-Pearson statistical hypothesis testing and consider whether it can be justified on the same grounds (3.1). We argue that it cannot be justified in this way because the purported nulls in our case studies lack the relevant features of statistical nulls (3.2). We then consider model selection theory as an alternative source of justification and argue that it is also insufficient to support the strategy in question (3.3). Finally, we explain why several additional ways of understanding pseudo-null hypotheses are also unsatisfactory (3.4).

3.1. Why Statistical Null Hypotheses are Treated Asymmetrically

Statistical hypothesis testing was developed in the 1920s and 30s most influentially by R. A. Fisher, Jerzy Neyman, and E. S. Pearson. Modern statistics has merged them and others into a toolbox of hybrid methods. A statistical hypothesis test is a procedure for making inferences about populations based on sample data taken from those populations. Fisher and Neyman-Pearson testing differ with respect to both the number of hypotheses being considered in a test and the possible outcomes of that test. In Fisher testing, one hypothesis (the null) is tested and the possible outcomes of the test are either rejection of this null or failure of rejection. In Neyman-Pearson testing, two hypotheses (the null and the alternative) are tested and the possible outcomes are either support for the alternative hypothesis (and rejection of the null) or retaining

of the null hypothesis (and no support for the alternative).¹

The strategy in our case studies are similar to Neyman-Pearson hypothesis testing in the number of hypotheses being tested and the asymmetrical treatment of hypotheses.² As Peter Godfrey-Smith puts it, the null hypothesis in Neyman-Pearson testing “typically gets the benefit of the doubt” (Godfrey-Smith 1994, 280). In this test, a type I error occurs when one rejects a true null hypothesis, while a type II error occurs when one retains a false null. Typically, fewer type I errors are allowed than type II errors: whereas the accepted rate of rejecting the null hypothesis when it is true (type I error) is at or below 5%, the accepted rate of retaining the null hypothesis when it is false (type II error) is usually 20% and can be up to 30% or 40% (Sani and Todman 2008). In this way, the null and alternative hypotheses are treated asymmetrically with the null hypothesis being more difficult to reject than the alternative.

This raises the question of how one determines which hypothesis should be treated as the null and what justifies its status. For Neyman and Pearson, the answer to this question was determined on pragmatic grounds. In discussing how one should decide which error to avoid, they write:

We are reminded of the old problem considered by Laplace of the number of votes in a court of judges that should be needed to convict a prisoner. Is it more serious to convict an innocent man or to acquit a guilty? That will depend upon the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be

¹ The language of “null hypothesis” comes from Fisher's single hypothesis testing procedure, and Neyman and Pearson objected to its use, but it has long been a part of Neyman-Pearson testing. See (Gigerenzer 2004).

² John Beatty (Beatty 1987) discusses the case of Kimura's neutral theory of molecular evolution as a null hypothesis with respect to Fisher's version of null hypothesis testing and explains why it is inappropriate.

controlled and minimised. (Neyman and Pearson 1933, 296)

Today, a more widely accepted justification for the asymmetrical treatment of error in Neyman-Pearson testing is that the null hypothesis is a hypothesis of “no effect” or “nothing going on.” Godfrey-Smith refers to this as a *semantic* justification because it concerns the content of the hypothesis, rather than the pragmatic implications of accepting it. Although Godfrey-Smith does not go into detail about what he takes the semantic justification to be, he holds it is grounded in the value of simplicity. He writes,

If hypotheses of “no effect” are nulls, then the asymmetry between alpha [the rate of type I errors] and beta [the rate of type II errors] operates as the wielder of Occam’s razor. The more serious error is multiplying effects beyond necessity, rather than not recognising enough effects. The asymmetry establishes a bias in favor of the simpler hypothesis.

(Godfrey-Smith 1994, 282)

We think that Godfrey-Smith is right here to point out that the null hypothesis is singled out in statistical hypothesis testing because it is taken to be a hypothesis of no effect, where “no effect” in an experimental context means that the independent variable being tested did not affect the dependent variable being measured. However, as we argue below, it is a mistake to understand this favoring as the wielding of Occam’s razor. Understanding the asymmetrical acceptance of error in statistical hypothesis testing requires situating it within the practices of experimental design and statistical inference.³

The role of the null hypothesis in statistical hypothesis testing is best understood within the context of a well-designed experiment.⁴ The purpose of an experiment is generally to test

³ For general introductions to statistical hypothesis testing see (Sani and Todman 2008, Dienes 2008).

⁴ We describe the statistical reasoning here in terms of experiments, but do not intend to exclude observational studies. When experiments are impractical, observational studies are done instead. In an experiment, participants are

whether one variable (the independent variable) has some effect on another (the dependent variable). Researchers do this by testing whether there is a difference in the dependent variable between conditions in which the independent variable is present (the experimental condition) and absent (the control condition). A well-designed experiment requires that one control all extraneous variables—i.e., those variables (other than the independent variable) that could or are known to influence the dependent variable. Researchers control these variables by either holding them constant (through repeated measures design, for example) or distributing them randomly across conditions (through the random allocation of participants, testing conditions, etc.). For example, the age and education level of a participant might affect how he or she performs on a memory task. Knowing this, researchers can randomly assign participants to the control and experimental conditions in order to randomly distribute these factors. Statistical inference is then used to judge whether these variables could have led to the observed experimental results.

Researchers use inferential statistics to determine the likelihood that randomized extraneous variables are responsible for the results of an experiment. A statistical null hypothesis is employed for this purpose. The statistical null hypothesis formalizes the situation in which there is no significant difference in the mean outcome of the control and experimental conditions, where a significant difference is one that can reasonably be attributed to the independent variable. If researchers could eliminate all extraneous variables, and the independent variable failed to affect the dependent variable, then the results of the control and experimental conditions would be identical. In such a case, if the independent variable produced no effect, there would be no other variable that could. A world without extraneous variables would be one

randomly assigned to the control or experimental group, while in an observational study, participants are selected by uncontrolled factors. Our account of statistical hypothesis testing extends to observational studies insofar as they employ the standard tools of statistical inference used in Neyman-Pearson testing. On the principles of designing an observational study and how to detect, minimize, and measure biases, see (Rosenbaum 2005).

in which factors like age or education had no effect on participants' performance on a memory task. In the real world, extraneous variables are always a source of variability. The question then is whether this variability is sufficient to account for the differences observed across conditions. If it is sufficient, then one accepts the null hypothesis that the observed difference is not significant—that is, not large enough to be attributed to the effects of the independent variable.

How does one determine whether an observed difference can be attributed to the independent variable? Roughly, one calculates a test statistic that reflects the variability of the data within conditions, where we know the independent variable is not having a differential effect. For example, researchers can estimate the effects variables like age and education have on participants' performance on a memory task by looking at how performance varies within the control condition and within the experimental condition. The variability within these conditions should be attributable to the randomized variables of age and education level and any other factors that have been randomized (socio-economic status, health, cultural background, etc.). This variability or standard deviation is then used to calculate the probability that the difference observed between conditions is due to extraneous variables alone. Again, if this probability is small (traditionally, less than 1 in 20), then one concludes that the observed difference is significant or the independent variable likely had an effect. If this probability is large, on the other hand, then one concludes that the observed results may be the product of the effects of randomized variables. In the latter case, the experiment has not produced evidence for a relationship between the independent and dependent variables.

Now we can see more clearly why researchers err on the side of accepting the null hypothesis in statistical hypothesis testing. It is not because it is simpler per se to posit no relationship between the independent and dependent variables. Instead, the null hypothesis is

avored because it represents the effects of extraneous variables that experimenters have intentionally tried to control. These effects are “known” in the sense that their quantity was determined by measuring the variability of the data within conditions. The purpose of the experiment then is to detect an effect on the dependent variable by the independent variable amidst the known noise or error created by the extraneous variables (or “nuisance variables” as they are also called). In other words, to retain the null hypothesis is to fail to rule out the possibility that the observed results were caused by those variables the researchers did their best to eliminate. In this way, the null hypothesis in statistical hypothesis testing is generally not the explanans of interest. In contrast, the effects of the independent variable are unknown and uncovering these effects is the purpose for which the experiment was designed. There may be other evidence suggesting that the independent variable should have an effect in this situation, but the aim of the study is to provide experimental evidence that this is the case.

Even in cases in which the statistical null hypothesis represents noise created by a natural phenomenon (as opposed to intentionally randomized variables), researchers generally do not seek to discover something about the null hypothesis, but rather use it as a tool for learning about the alternative hypothesis. When researchers announced the discovery of the Higgs boson, for example, they did so on the grounds that there was a tiny chance (1 in 3.5 million) that the observed decay patterns were the result of background processes alone (Staley 2017). If, in contrast, the observed patterns were such that there was a high probability that they were produced by these background processes, this would be interpreted as a lack of evidence for the Higgs boson, rather than evidence in support of the existence of background processes. The investigative question was not whether background processes could produce decay patterns resembling those of the hypothesized Higgs boson—researchers knew this was the case, hence

the need to control for it—but rather whether Higgs bosons contributed to the observed effects. Statistical nulls are generally not claims about some feature of the world in need of explanation. Instead, they are claims about a feature of the world investigators have tried their best to control in order to learn something about an alternative, experimental hypothesis.

Given the above, retaining the null hypothesis is often considered a negative result. The contrary view, that the failure to reject the null hypothesis is in itself positive evidence of no effect, commits the “fallacy of acceptance” (Mayo and Spanos 2006, 338). Although it could be that the independent variable had no effect on the dependent variable and that the observed data reflects the actions of extraneous variables alone, it could also be that the independent variable had an effect, but the experiment was unable to detect it due to inadequate power, poor measurement, experimenter error, poor sampling, or various other factors. One way of eliminating some of these alternative explanations is to increase the statistical power of one’s test. Statistical power is $1 - \beta$, where β is the probability of a type II error or the error of retaining the null hypothesis when it is false. The power of a test then is the probability of correctly rejecting a false null. It is recommended that the power of a test be set to 0.80 ($\beta = 0.20$), but in practice few researchers conduct power analyses and the power of experiments in the behavioral sciences is notoriously low (Sedlmeier and Gigerenzer 1989, Cohen 1988). The lower the power of a test, the less likely one will detect an effect when one is in fact present. Thus, in practice, retaining the null is usually taken to be a negative result.

Even if one were to conduct a test with very high power (with large sample sizes, for example), it remains that statistical nulls are often constructed in such a way that it does not make sense to seek evidence in their favor. As mentioned above, typically these hypotheses represent factors that are known or uninteresting such as the effects of nuisance variables or

other sources of noise. Additionally, many statistical nulls are strictly false. Null hypotheses are often point hypotheses, which are useful for constructing a well-defined sampling distribution, but researchers do not expect to find such exact relationships and effects in nature (e.g., that the difference between the mean scores of two populations is exactly zero). The alternative hypothesis, on the other hand, tends to be non-specific (such as the means will differ by some unspecified amount). If the alternative hypothesis were tested directly, a sampling distribution would be needed for each difference the experimenter wanted to test. It is more practical to test the null hypothesis.

Statistical null hypotheses serve a particular role in experimental practice. By taking into account those causes that researchers know are having an effect on a dependent variable, one can more accurately determine whether the independent variable under investigation is having an effect (the alternative hypothesis) or whether the difference observed across conditions is attributable to the effects of extraneous variables alone (the null hypothesis). Although the bar for retaining the null hypothesis is lower than the bar for accepting the alternative hypothesis, the null hypothesis is generally not epistemically privileged because researchers take retaining the null as providing little to no new information about the causes and effects posited by the null hypothesis. If the power of an experiment is low, and the null hypothesis is retained, this provides no new information, as the chances of detecting an effect by the independent variable were small to begin with, so failing to detect such an effect is uninformative. If the power of an experiment is high, and other sources of error are unlikely, then one might be able to conclude that the observed effects are due to those causes posited by the null hypothesis; however, this provides more information about the alternative hypothesis (the independent variable had no effect in this situation) than the null hypothesis because the effects of the null hypothesis were

already known (as measured by within-condition variability) and the causes of those effects were assumed to begin with.

The pseudo-null strategy is analogous to Neyman-Pearson hypothesis testing in several respects. It relies on two hypotheses, a null and an alternative; it couples rejection of one hypothesis with the retainment of the other; and it treats the hypotheses asymmetrically. Given these similarities, perhaps this strategy is justified as an extension of statistical null hypothesis testing. In the next section, we argue that this is not the case.

3.2. Why Pseudo-Nulls are Not Statistical Nulls

The role statistical nulls play in hypothesis testing depends on the context of experimental design and statistical inference in which they are used. The fact that researchers require that one reject the null before accepting the alternative hypothesis makes sense within this context. Accepting the null means the noise created by randomized extraneous variables was too large for the potential effects of the independent variable to be detected. In practice, accepting the null is a negative finding about the alternative, experimental hypothesis.

Proponents of neutral theory and behavior-reading refer to these hypotheses as “nulls” and employ the same inferential strategy as found in statistical hypothesis testing: they hold that a necessary condition for having evidence for the alternative hypothesis is rejecting the null and that if one fails to reject the null, it should be accepted as the best explanation for the phenomenon under investigation. In this section, we show that this strategy lacks the relevant features of statistical null hypothesis testing and thus cannot be justified on the same grounds. Our point is not that the behavior-reading hypothesis and neutral theory cannot be used to supply appropriate null hypotheses in any line of reasoning, but that they do not supply appropriate null

hypotheses in the reasoning strategies actually being employed.

There are several important disanalogies between the neutral theory and behavior-reading hypothesis, on the one hand, and statistical null hypotheses, on the other. Perhaps most critically, the former are not used in statistical inference. In comparative psychology, many experiments have been conducted to test for mindreading in animals (Call and Tomasello 2008). The statistical null hypothesis in these cases is that the observed effect is due to randomized extraneous variables. The behavior-reading hypothesis does not feature as a statistical null in these experiments, but rather as a competing hypothesis concerning the cognitive mechanism responsible for the observed behaviors. Indeed, the behavior-reading hypothesis does not make predictions about the distribution of data in a given experiment. When it does speak to the outcome of an experiment, it does so in an ad hoc manner, “predicting” after the fact the very outcome expected by the alternative hypothesis (Fletcher and Carruthers 2013; Halina 2015). In other words, the behavior-reading hypothesis typically takes the positive results of mindreading experiments (long after statistical analysis has taken place) and provides alternative causal explanations for these results.

Community ecologists have sampled communities from all over the world and used the observed distributions to test whether ecological selection is operating. The appropriate statistical null hypothesis for such tests are that the observed species distributions will not differ significantly from what we would expect if individuals of different species were distributed at random in space and time (Bell 2000, 614), where what is meant by “random” can be filled in with different statistical measures. Such statistical nulls are silent about what might cause a particular distribution; instead, they specify the statistical distribution against which the data should be compared. In contrast, neutral theory makes a claim about the causes of species

distributions (that they are the result of drift, immigration, and speciation), without holding that those distributions themselves will be random in any quantitative or statistical sense (Bell 2000).

Hubbell implicitly has this kind of distinction in mind when he first introduces “neutrality”,

Before proceeding, I need to be more precise about the meaning of *neutrality* as used in this book. Despite its moniker, the concept of neutrality actually has many meanings in the literature. To most people, the word *neutral* congers [*sic*] up the qualitative notion of “nothing going on.” But exactly what people mean by this phrase often turns out to differ from one person to the next. I use *neutral* to describe the assumption of per capita ecological equivalence of all individuals in a trophically defined community. This is a very unrestrictive and permissive definition of *neutrality* because it does not preclude interesting biology from happening or complex ecological interactions from taking place among individuals. (Hubbell 2001, 6-7)

Neutral and selection theories include interesting biology with complex interactions and offer competing explanations for the observed abundance distribution once the statistical null hypothesis has been rejected.

Neither the behavior-reading hypothesis nor neutral theory play the role of statistical nulls in experiments in comparative psychology and community ecology. The experimental and statistical machinery which grounds the logic of rejecting the null hypothesis as a condition for finding support for the alternative is missing in the case of these pseudo-nulls.

Proponents of neutral theory and behavior-reading might object that, although these hypotheses do not feature in statistical inference, the causal variables that they posit are either (i) presupposed by the alternative hypotheses--selection mechanisms and mindreading--or (ii)

“known” in the sense of already thought to be operating in the situations under investigation. However, neither of these are the case.

Let us begin with the comparative psychology case. Concerning (i), the ability to mindread does not presuppose the behavior-reading hypothesis. The behavior-reading hypothesis is taken by both sides of the debate to be an independent and competing causal explanation for the success of subjects on mindreading tasks. It holds that animals possess precisely those behavioral-rules required to succeed on a given mindreading task. According to proponents of the behavior-reading hypothesis, the mindreading hypothesis makes a different causal claim. Penn and Povinelli characterize the difference between nonhuman behavior-reading and human mindreading as follows:

we believe that human and nonhuman animals possess a variety of mechanisms for recognizing those relations that are causally relevant to predicting the goal-directed behavior of other intentional agents. These heuristics enable both human and nonhuman animals to pick out the causally relevant relations between ‘what’ an agent is ‘looking’ at and how that agent is likely to behave in the near future... However, only humans cognize the higher-order analogical similarities between perceptually disparate behaviors and thus only humans possess the ability to reinterpret other agent’s goal-directed relations in terms of abstract mental state relations disembodied from any particular task context. (Penn and Povinelli 2013, 20)

The hypothesis that nonhuman animals behavior-read is not merely the claim that they lack the ability to represent the mental states of others, but also that they possess a set of heuristics that allows them to make the correct behavioral predictions without these abilities. If the human ability to reinterpret behavior in terms of mental states had no causal effects, then this ability

would be superfluous and impossible to test empirically. Povinelli and colleagues, however, do not think this is the case (Penn and Povinelli 2007).

Concerning (ii), although researchers might hold that behavior-reading is used by some animals in some contexts, whether it is this process that leads to the observed behavior under investigation in a given experiment, as opposed to mindreading, is precisely the question in dispute. Researchers might hold that, for instance, great apes are capable of behavior-reading where behavior-reading is understood as the general ability to learn associations between environmental cues and behaviors. But it is a different question whether apes possess precisely those set of behavioral rules required to succeed on this or that false belief task, for example. Currently, not only is there no consensus that apes have such rules (with respect to most mindreading experiments), but many researchers find this hypothesis implausible (Call and Tomasello 2008, Halina 2015; Krupenye et al. 2016). This does not mean that evidence for this hypothesis will never be forthcoming, but rather that no convincing evidence has been presented to date. Indeed, proponents of the behavior-reading hypothesis have generally not attempted to present such evidence precisely because they take their hypothesis to be the null. This, however, begs the question. Contrast the behavior-reading hypothesis with the statistical null hypotheses actually used in mindreading experiments. These concern variables such as trial order, subject allocation, food allocation, the location of occluders, and so forth--variables that are randomised or held constant across experimental and control conditions and are uncontested regarding their presence and effects in the context under investigation.

In community ecology, regarding (i), drift, immigration, and speciation operate independently of selection and are not formally presupposed by selection. In virtue of assuming neutrality, neutral theory excludes selection. But neutrality leaves open which processes are

responsible for the patterns of diversity and abundance of species seen within communities. In population level and individual-based models, those that include selection and operate within a finite population do necessarily include drift. However, drift is arguably due to finite population size. In such models with infinite population size, selection occurs without drift.⁵ In other kinds of models, many of them include a specific form of selection without drift (Vellend 2016, 63-66). Further, neither selection nor drift requires immigration or speciation to operate and many models of selection include neither immigration nor speciation.

Regarding (ii), drift, immigration, and speciation are widely accepted as being causally relevant to abundance patterns across ecological communities. However, this does not imply that all three processes are influencing a particular community at a given time or that selection is not occurring. In fact, ecologists sometimes begin by assuming selection as the main cause of patterns of diversity and abundance within a community. For example, David Tilman has spent almost 40 years developing and applying Resource-Ratio or R^* Theory, a selection theory of communities based on only the competition of species for limited resources (Tilman 1982). Tilman holds that this theory is useful to "see what features of a system can and cannot be explained using those processes" (Tilman 1987, 135). The many other available selection-based theories and models of communities can be used in the same way. Drift, immigration, and speciation are neither more nor less known to be general causes of abundance patterns than selection (Vellend 2016). The question facing ecologists is which subset of these four processes needs to be invoked, and to what relative degree, to explain a given pattern (Beatty 1997).

⁵ This is controversial and depends in part on how selection and drift are conceptualized. See (Millstein 2017, §2) for an overview of the fault lines.

3.3. What about Model Selection Theory?

We have argued that pseudo-null hypotheses like the neutral theory and behavioral-reading hypothesis are relevantly disanalogous from statistical null hypotheses. Given this, they should not receive the epistemic privileges they currently enjoy. One might object to this by arguing that this epistemic privilege comes not from statistical null hypothesis testing, but from model selection theory.⁶ In this section, we argue that this is not the case.

A statistical model differs from a statistical hypothesis in having free parameters. The linear equation $y=2x+3$ is a statistical hypothesis. Determining the probability of some data given this hypothesis is straightforward because you can measure the fit of the line to the data via their difference. The linear equation $y=mx+b$, where m and b are parameters free to vary, is a statistical model. The difference between this equation and some data is not well-defined. But sometimes one wants to determine if the data are more linear or quadratic. Model selection criteria can help to answer this.

The Akaike information criteria gives a uniquely best way of balancing the respective virtues of the fit between a model and data set and the number of free parameters in that model (Forster and Sober 1994).⁷ A model's number of free parameters is considered the measure of its simplicity in the Akaike framework. Hirotugu Akaike proved a mathematical theorem that determines whether the additional cost of added complexity in terms of the number of free parameters is worth the benefit of increased fit. Given a set of statistical models and a sampled data set, the Akaike criterion selects the best model, defined as the model which could have generated the data set and is most likely to be predictively accurate with respect to future data

⁶ We thank Elliott Sober for raising this point.

⁷ Other model selection criteria exist, such as the Bayesian information criteria. While there are important differences between these selection criteria, our arguments here extend to them as well.

drawn from that data set. The Akaike criterion rewards models for their goodness-of-fit to the given data set and penalizes them for their number of free parameters. With too much goodness-of-fit or too few free parameters, the model is over-fit and more sampled data will probably make the fit worse. Akaike's theorem gives a unique way of weighing a model's goodness-of-fit against its number of free parameters. The model with the smallest value is selected. In this way models are rewarded for their simplicity.

Simplicity is traditionally considered to be a theoretical virtue, distinct from empirical virtues concerned with evidence. But Elliott Sober argues that the choice of a best model using the Akaike criterion is an empirical decision in the same way a test using only goodness-of-fit is empirical. He writes that, "although it is clear that simplicity is a separate consideration in model selection from fit to data, the justification provided by Akaike's theorem for using simplicity depends on empirical assumptions. Simplicity is therefore an empirical consideration" (Sober 2002, S117). Given this, perhaps the pseudo-nulls in our case studies should be privileged because their simplicity makes them more empirically successful than the alternatives. Although it may seem that the two hypotheses fit the data equally well, the pseudo-nulls are superior because they have fewer free parameters. Simplicity trumps the theoretical virtues held by the alternative hypothesis, such as coherence with other theories or generality, in the same way goodness-of-fit trumps these theoretical virtues, or so the story goes.

We have three objections to this strategy. First, the Akaike framework selects the best statistical model from a limited set. But because the set of models is not exhaustive, the best model may just be the best of a bad lot. There is no further assessment of how good the best model is at fitting the data. This is a general weakness of applying the Akaike framework to scientific models.

Second, the pseudo-nulls in our case studies are viewed as simpler because they purport to exclude a cause included in the alternative hypotheses. Neutral theory is simpler than selection theory because it does not posit selection acting in a community; behavior-reading is simpler than mindreading because it does not posit that individuals represent mental states. However, simplicity qua free parameters is not the same as simplicity qua excluding a cause (Sober 2009, 247). Free parameters are a feature of mathematical models and causes are features of theories, hypotheses, and interpretations of mathematical models. Two models might be interpreted in terms of the same causes and differ in their number of adjustable parameters—treating two causes as interactive rather than additive, for example, adds a free parameter. One model might also be interpreted as including fewer causes than another while having more free parameters. For example, the broken-stick model is a selection model with two free parameters: number of species and number of individuals (MacArthur 1957). However, the neutral model has three free parameters: local community size, immigration rate, and a function of metacommunity size and speciation rate. The neutral model is simpler than the broken-stick model in the sense of excluding selection as a cause of species abundance distributions, but not in the sense required by the Akaike framework. The reasoning strategy that casts the hypotheses in our case studies as nulls cannot be defended using model selection theory without first showing that they have fewer adjustable parameters than their competitors.

Third, even if proponents of pseudo-nulls address the previous two objections, the reasoning strategy itself is not suited to justification from a formal model selection framework. The Akaike criterion is a tool for comparing mathematical models on two dimensions: fit with a data set and number of free parameters. For identifying and evaluating the mechanisms responsible for a given distribution of data, the Akaike criterion is silent.

The mismatch between the Akaike criterion and the criteria needed to evaluate mechanisms becomes clear in Purves and Pacala's critique of the neutral theory. They constructed a selection model that was mathematically equivalent to a formalization of the neutral theory and therefore equally well supported (Purves and Pacala 2005). They argued neutral patterns do not imply neutral processes by showing that a community under strong selection can look just the same as under pure neutral theory. They imagine a metacommunity made up of two local communities so different from each other that the members of either community cannot live in the other. Let these subcommunities operate according to neutral theory independently of each other. Purves and Pacala prove that, because the patterns of the subcommunities are described accurately by neutral theory, those of the metacommunity will be as well. Their proof is an instance of a larger issue. Within certain limits, neutral theory accurately describes communities in which both (a) every individual is functionally equivalent and so described by the same parameter value, and (b) every individual is functionally unique but the average value of the individuals' parameter values is used. Hubbell acknowledged the point that neutral patterns underdetermine neutral processes, yet he argued that the two explanations were not on a par. This is the context of his conclusion, quoted above, that "obtaining acceptable fits from neutral models shifts the burden of proof to those who would assert that more complex theory is required to explain nature" (Hubbell 2006, 1387). Hubbell could not be referring to the number of free parameters in the two models here because the two models are identical. Only their causal interpretations are distinct. Hubbell means that the neutral theory should be favored on the grounds of its relative simplicity qua excluding selection. If proponents of pseudo-nulls move in the direction required to apply model selection theory, they may well lose their simpler status because such models do not in general have fewer free parameters.

3.4 Other Potential Sources of Justification

The pseudo-null strategy may have other sources of justification besides statistical hypothesis testing and model selection theory. However, we think these readings are the most natural and believe most readers will interpret “null” language in this way. The term “null hypothesis” cannot easily be distanced from these contexts. These justifications are also the strongest rhetorically because proponents of the strategy need only use the term “null hypothesis” to suggest a solid methodology for their reasoning. If proponents of pseudo-nulls intend a different reading of their inference strategy, then they must make this clear. Our point in this paper is that the obvious readings are unjustified and any others are left implicit. However, because a reader may suspect there are additional plausible sources of justification, we briefly explain why we find several unsatisfactory.

First, “null hypothesis” could mean the simpler hypothesis, where simpler hypotheses have Occam's Razor on their side. If this is the case, then there are good reasons to think that the reasoning will remain unjustified. Many philosophers have critiqued the use of simplicity both in the case of comparative psychology and general scientific reasoning (Dacey 2016, Fitzpatrick 2008, 2017, Meketa 2014, Sober 2005).

Second, “null hypothesis” could mean *null model* as used by ecologists (Gotelli and Graves 1996). Null modeling tests a hypothesis that a set of processes is causally responsible for a set of patterns by comparing the data with a model which excludes those processes (Bausman 2018). If this is the case, then the reasoning strategy is still unjustified because in null modeling only the alternative model being tested by the null model is able to gain evidence.

Third, “null hypothesis” could mean *baseline model* as used by ecologists. Baseline

modeling measures the relative significance of various processes responsible for a token instance of a pattern by comparing the data with a model that includes only the baseline processes. Any deviations from the baseline expectation are taken to be caused by additional processes (Bausman 2018). If this is the case, however, proponents of the behavior-reading and neutrality hypotheses need to justify taking their processes as the baseline.

Fourth, “null hypothesis” could mean *default* hypothesis, where the defaults are given the benefit of the doubt for some special quality they possess. If this is the case, then we ask that the details of this strategy be specified and distanced from formal methods.

Finally, perhaps “null hypothesis” marks the hypothesis as having some higher empirical evidence or Bayesian prior probability. If this is the case, then this is an evidence-based argument and not based on *prima facie* reasons. The association to formal methodologies, theoretical virtues, and indeed falsificationism needs to be dropped.

We want the reasoning strategy to be made plain and explicit both so that it can be critiqued and so that it can be understood. When scientists cloak their arguments in misleading language, it gives the impression both that outdated methodological ideas such as Occam's Razor and falsificationism are alive and well in science and that overpowered and not-well-understood formal methodologies are necessary.

4. Parity Please

We have argued that the epistemic privileging of pseudo-nulls is not justified as a form of statistical null hypothesis testing. Like selection theory and mindreading, neutral theory and the behavior-reading hypothesis are not statistical hypotheses and are not being used as such. The appeal to these latter hypotheses as “nulls” is a rhetorical strategy without any clear source of

justification. In our view, this strategy presumes that the virtue of simplicity trumps all others without providing good reasons for it doing so. Using the term “null hypothesis” serves to gloss over this in the name of good scientific, statistical methodology.

Richard Levins made the tradeoffs between scientific virtues among different models and modeling practices well known (Levins 1966). He argued that simplicity is one virtue among many that all exist in complex tradeoff relationships. Which virtues should be valued over others depends on the problems, goals, and circumstances of the particular situation. We find such tradeoffs in our case studies. For example, species similarities and differences are used to explain a wide variety of empirical results, such as why a particular type of tree is found in the same temperature belt across continents. This result is not in the neutral theory’s explanatory scope. Appealing to the neutral theory on the grounds of relative simplicity in order to explain some patterns of biodiversity and biogeography means researchers are choosing to value simplicity over generality. Instead of relative species abundance distributions and continental species patterns being explained in the same way, they would be explained with different theories. This may be a tradeoff researchers endorse, but it is a choice. Conceiving neutral theory as a null takes this choice away from researchers. It makes the favoring of simplicity uncontroversial and inevitable. Tradeoffs between virtues require choices, which in turn require reasons and justification that can be discussed, critiqued, and evaluated.⁸

In addition to being weighed against other virtues, the value of simplicity must be assessed against the particular context in which it is employed. Sober (2002), John Norton (2003), and Helen Longino (2008) have argued that simplicity is not a general virtue, but is, whenever justified, shorthand for some set of facts which make the simpler hypothesis more

⁸ Such a discussion between simplicity and other scientific virtues may be locally (Levins 1966) or generally (Douglas 2009) justified.

likely. As Sober writes,

When a scientist uses the idea [of parsimony], it has meaning only because it is embedded in a very specific context of inquiry. Only because of a set of background assumptions does parsimony connect with plausibility in a particular research problem.

What makes parsimony reasonable in one context therefore may have nothing in common with why it matters in another. (Sober 1994, 140)

If these authors are right, then an appeal to simplicity must be evaluated on a case-by-case basis.

It may be that simplicity should trump other virtues, in the way that empirical adequacy and goodness-of-fit sometimes does.⁹ However, until a good argument is provided for this, the value of simplicity should be treated on a par with other virtues. The pseudo-null strategy is an attempt to move hypotheses away from parity by shifting the burden of disproving the null to the alternative hypotheses on the authority of statistics. As we have argued, there is no clear justification for this strategy, however, so the hypotheses should be treated on a par.

5. Conclusion

The pseudo-null strategy is employed on behalf of the neutral theory of ecology and behavior-reading hypothesis in comparative psychology. Because the strategy presents a rhetorically powerful solution to a common problem, we expect these are not isolated cases. However, we have argued that this strategy is not justified as an extension of statistical inference. Hypotheses should not be privileged as “nulls” unless a valid justification has been given for

⁹ We have focused on simplicity because it is directly appealed to by the proponents of the behavior-reading and neutrality hypotheses to justify privileging them. But we can imagine another virtue being privileged at the cost of all others, generality for example, and we would be against its use as a way of treating two hypotheses asymmetrically and calling the more general one the ‘null’ just the same.

such privileging.

References

- Bausman, William C. 2018. "Modeling: Neutral, Null, and Baseline." *Philosophy of Science*. 84(4).
- Beatty, John. 1987. "Natural Selection and the Null Hypothesis." In *The Latest on the Best*, edited by John Dupre. Cambridge: MIT Press
- Beatty, John. 1997. "Why Do Biologists Argue like They Do?" *Philosophy of Science* 64 (Proceedings):S432-S443.
- Bell, Graham. 2000. "The Distribution of Abundance in Neutral Communities." *The American Naturalist* 155 (5):606-617.
- Call, Josep, and Michael Tomasello. 2008. "Does the Chimpanzee have a Theory of Mind? 30 Years Later." *Trends in cognitive sciences* 12 (5):187-192.
- Chase, Jonathan M, and Mathew A Leibold. 2003. *Ecological Niches: Linking Classical and Contemporary Approaches*. Chicago: University of Chicago Press.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge Academic.
- Dacey, Mike. 2016. "The Varieties of Parsimony in Psychology." *Mind & Language* 31 (4):414-437.
- Dienes, Zoltan. 2008. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. New York: Palgrave Macmillan.
- Douglas, Heather. 2009. *Science, Policy, and The Value-Free Ideal*: University of Pittsburgh Pre.

- Fitzpatrick, Simon. 2008. "Doing Away with Morgan's Canon." *Mind & Language* 23 (2):224-246.
- Fitzpatrick, Simon. 2017. "Against Morgan's Canon." In *The Routledge Handbook of Philosophy of Animal Minds* edited by Kristin Andrews and Jacob Beck. Routledge.
- Fletcher, Logan, and Peter Carruthers. 2013. "Behavior-Reading versus Mentalizing in Animals." In *Agency and Joint Attention*, 82-99. Oxford: Oxford University Press.
- Forster, Malcolm, and Elliott Sober. 1994. "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science* 45 (1):1-35.
- Gigerenzer, Gerd. 2004. "Mindless Statistics." *The Journal of Socio-Economics* 33 (5):587-606.
- Godfrey-Smith, Peter. 1994. "Of Nulls and Norms." PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association.
- Gotelli, Nicholas J, and Gary R Graves. 1996. *Null Models in Ecology*. Washington DC: Smithsonian Institution Press.
- Halina, Marta. 2015. "There Is No Special Problem of Mindreading in Nonhuman Animals." *Philosophy of Science* 82 (3):473-490.
- Hare, Brian, Josep Call, and Michael Tomasello. 2006. "Chimpanzees Deceive a Human Competitor by Hiding." *Cognition* 101 (3):495-514.
- Hubbell, Stephen P. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography* Edited by Simon A. Levin and Henry S. Horn, *Monographs in Population Biology*. Princeton: Princeton University Press.
- Hubbell, Stephen P. 2006. "Neutral Theory and the Evolution of Ecological Equivalence." *Ecology* 87 (6):1387-1398.

- Krupenye, Christopher, Fumihiro Kano, Satoshi Hirata, Josep Call, and Michael Tomasello. 2016. "Great apes anticipate that other individuals will act according to false beliefs." *Science* 354 (6308):110-114.
- Levins, Richard. 1966. "The Strategy of Model Building in Population Biology." *American Scientist*:421-431.
- Lloyd, Elisabeth A. 2015. "Adaptationism and the Logic of Research Questions: How to Think Clearly About Evolutionary Causes." *Biological Theory*:1-20.
- Longino, Helen E. 2008. "Values, Heuristics, and The Politics of Knowledge." In *The Challenge of the Social and the Pressure of Practice: Science and Values Revisited*, edited by Martin Carrier, Don Howard and Janet Kourany, 68-86. Pittsburgh: University of Pittsburgh Press.
- Lurz, Robert W. 2011. *Mindreading Animals: The Debate Over What Animals Know About Other Minds*. Cambridge: MIT Press.
- MacArthur, Robert H. 1957. "On The Relative Abundance of Bird Species." *Proceedings of the National Academy of Sciences of the United States of America* 43 (3):293.
- Mayo, Deborah G, and Aris Spanos. 2006. "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction." *The British Journal for the Philosophy of Science* 57 (2):323-357.
- Meketa, Irina. 2014. "A Critique of the Principle of Cognitive Simplicity in Comparative Cognition." *Biology & Philosophy* 29 (5):731-745.
- Melis, Alicia P, Josep Call, and Michael Tomasello. 2006. "Chimpanzees (Pan troglodytes) Conceal Visual and Auditory Information from Others." *Journal of Comparative Psychology* 120 (2):154.

- Millstein, Roberta L. 2017. Genetic Drift. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.
- Neyman, Jerzy, and Egon S Pearson. 1933. "The Testing of Statistical Hypotheses in Relation to Probabilities A Priori." *Mathematical Proceedings of the Cambridge Philosophical Society*.
- Norton, John D. 2003. "A Material Theory of Induction." *Philosophy of Science* 70 (4):647-670.
- Penn, Derek C, Keith J Holyoak, and Daniel J Povinelli. 2008. "Darwin's Mistake: Explaining the Discontinuity Between Human and Nonhuman Minds." *Behavioral and Brain Sciences* 31 (02):109-130.
- Penn, Derek C, and Daniel J Povinelli. 2007. "On the Lack of Evidence that Non-Human Animals Possess Anything Remotely Resembling a 'Theory of Mind'." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362 (1480):731-744.
- Penn, Derek C, and Daniel J Povinelli. 2009. "On Becoming Approximately Rational: The Relational Reinterpretation Hypothesis." In *Rational animals, irrational humans*, edited by Shigeru Watanabe, Aaron P. Blaisdell, Ludwig Huber and Allan Young, 23-43. Tokyo: Keio University Press.
- Penn, Derek C, and Daniel J Povinelli. 2013. "The Comparative Delusion: The "Behavioristic/Mentalistic" Dichotomy in Comparative Theory of Mind Research." In *Agency and Joint Attention*, edited by Janet Metcalfe and Herbert S Terrace, 62-82. Oxford: Oxford University Press.
- Povinelli, Daniel J, and Jennifer Vonk. 2004. "We Don't Need a Microscope to Explore the Chimpanzee's Mind." *Mind & Language* 19 (1):1-28.

- Purves, Drew W, and Stephen W Pacala. 2005. "Ecological Drift in Niche-Structured Communities: Neutral Pattern Does Not Imply Neutral Process." In *Biotic Interactions in the Tropics: Their Role in the Maintenance of Species Diversity*, edited by David Burslem, Michelle Pinard and Sue Hartley, 107-138. Cambridge: Cambridge University Press.
- Rosenbaum, Paul R. 2005. Observational Study. In *Encyclopedia of Statistics in Behavioral Science*, edited by Brian S Everitt and David C Howell. Chichester: John Wiley & Sons.
- Sani, Fabio, and John Todman. 2008. *Experimental Design and Statistics for Psychology: A First Course*: John Wiley & Sons.
- Sedlmeier, Peter, and Gerd Gigerenzer. 1989. "Do Studies of Statistical Power Have an Effect on the Power of Studies?" *Psychological Bulletin* 105 (2):309.
- Sober, Elliott. 2005. "Comparative Psychology Meets Evolutionary Biology: Morgan's Canon and Cladistic Parsimony." In *Thinking with Animals: New Perspectives on Anthropomorphism*, edited by Lorraine Daston and Gregg Mitman, 85-99. Columbia University Press.
- Sober, Elliott. 1994. "Let's Razor Ockham's Razor." In *From a Biological Point of View: Essays in Evolutionary Philosophy*, 136-157. Cambridge University Press.
- Sober, Elliott. 2002. "Instrumentalism, Parsimony, and the Akaike Framework." *Philosophy of Science* 69 (S3):S112-S123.
- Sober, Elliott. 2009. "Parsimony and Models of Animal Minds." In *The Philosophy of Animal Minds*, edited by Robert W Lurz, 237. Cambridge: Cambridge University Press.
- Staley, Kent W. 2017. "Pragmatic Warrant for Frequentist Statistical Practice: The Case of High Energy Physics." *Synthese* 194 (2):355-376.

Tilman, David. 1986. "Resources, Competition and the Dynamics of Plant Communities." In *Plant ecology*, edited by M. Crawley, 51-75. Oxford, England: Blackwell Scientific Publications.

Tilman, David. 1982. *Resource Competition and Community Structure*. Edited by Simon A. Levin and Henry S. Horn, *Monographs in Population Biology*. Princeton: Princeton University Press.

Vellend, Mark. 2016. *The Theory of Ecological Communities*. Edited by Simon A. Levin and Henry S. Horn, *Monographs in Population Biology*. Princeton: Princeton University Press.